The Modern Language Journal

Do Explicit Instruction and High Variability Phonetic Training Improve Nonnative Speakers' Mandarin Tone Productions?

SETH WIENER,[1] MARJORIE K. M. CHAN,[2] and KIWAKO ITO[3]
[1] *Carnegie Mellon University, Department of Modern Languages, 160 Baker Hall, 5000 Forbes Ave., Pittsburgh, PA, 15213 Email: sethw1@cmu.edu*
[2] *The Ohio State University, Department of East Asian Languages and Literatures, Hagerty Hall, 1775 College Road, Columbus, OH 43210 Email: chan.9@osu.edu*
[3] *University of Newcastle, School of Humanities and Social Sciences, University Drive, Callaghan, NSW 2308, Australia Email: kiwako.ito@newcastle.edu.au*

This study examines the putative benefits of explicit phonetic instruction, high variability phonetic training, and their effects on adult nonnative speakers' Mandarin tone productions. Monolingual first language (L1) English speakers ($n = 80$), intermediate second language (L2) Mandarin learners ($n = 40$), and L1 Mandarin speakers ($n = 40$) took part in a multiday Mandarin-like artificial language learning task. Participants were asked to repeat a syllable–tone combination immediately after hearing it. Half of all participants were exposed to speech from 1 talker (low variability) while the other half heard speech from 4 talkers (high variability). Half of the L1 English participants were given daily explicit instruction on Mandarin tone contours, while the other half were not. Tone accuracy was measured by L1 Mandarin raters ($n = 104$) who classified productions according to their perceived tonal category. Explicit instruction of tone contours facilitated L1 English participants' production of rising and falling tone contours. High variability input alone had no main effect on participants' productions but interacted with explicit instruction to improve participants' productions of high-level tone contours. These results motivate an L2 tone production training approach that consists of explicit tone instruction followed by gradual exposure to more variable speech.

Keywords: Mandarin Chinese; lexical tone; speech production; explicit instruction; phonetic variability; second language acquisition

LANGUAGES USE DIFFERENT SPEECH sounds to convey meaning. Spoken communication in a second language (L2) therefore requires learning and implementing new articulatory patterns and producing novel speech sounds (Kormos, 2014; Laver, 1994). For most adults, accurate L2 speech production is notoriously difficult, and it can take years to achieve native-like abilities (e.g., Derwing & Munro, 1997, 2005; Flege, Munro, & MacKay, 1995; Hao, 2012; Munro & Derwing, 1995; Trofimovich & Baker, 2006). Given these challenges, a central focus in L2 pronunciation research concerns how to facilitate improvement in learners' speech through targeted instruction (see Lee, Jang, & Plonsky, 2014, for a recent review). In this study, we evaluate two approaches to improving L2 speech: explicit phonetic instruction (e.g., Saito, 2013; Saito & Wu, 2014) and high variability phonetic training (e.g., Thomson, 2012). As we outline in the subsequent sections, the benefits of these approaches and their implications for L2 classroom teaching of a tonal language remain unclear.

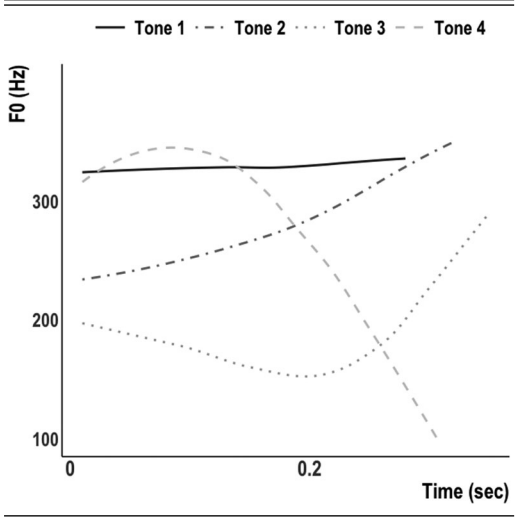
The Modern Language Journal, 0, 0, (2020)
DOI: 10.1111/modl.12619
0026-7902/20/1–17 $1.50/0

MANDARIN LEXICAL TONE

Mandarin Chinese serves as our target L2. Whereas L2 Mandarin programs and enrollment at the university level in the United States have grown considerably in the last two decades (ACTFL, 2012; Goldberg, Looney, & Lusin, 2015; Li, Wen, & Xie, 2014), this increased demand has not yielded great numbers of advanced L2 Mandarin learners as measured by proficiency test results (Cai, Chen, & Wang, 2010; Everson & Shen, 2010; Wu & Ortega, 2013). One potential reason for the dearth of proficient L2 Mandarin speakers in the United States is the well-documented tone perception and production challenges facing L1 English speakers (Hao, 2012, 2018a, 2018b)—challenges which often cause frustrated learners to abandon classroom learning before reaching a high level of L2 proficiency (see Ke, 2018; Pelzl, 2019; Zhang, 2018, for recent overviews). As an example, recent corpus evidence from over 300 beginner L2 Mandarin speakers indicates that roughly a third of all nonnative utterances are produced with a tonal error (Chen et al., 2016).

In terms of their acoustics, the four Mandarin tones are primarily characterized by their fundamental frequency (F0) contours (Gandour, 1983; Ho, 1976; Howie, 1976; Tseng, 1981), though secondary cues include duration, amplitude, and F0 turning point (Blicher, Diehl, & Cohen, 1990; Moore & Jongman, 1997; Shen & Lin, 1991; Whalen & Xu, 1992). Figure 1 plots the four tones in isolation as produced by a native female speaker from Beijing.

FIGURE 1
Mandarin Tones Produced in Isolation by a Female Native Speaker From Beijing



Tone 1 is produced with a high-level F0 and a relatively short duration. Because speakers must primarily attend to only F0 height, and because L1 English speakers are initially more sensitive to level tones than contour tones when first exposed to Mandarin (Huang & Johnson, 2010), L2 learners tend to acquire Tone 1 earlier than the other tones (Hao, 2012; Yang, 2015) and produce Tone 1 with relatively high accuracy (Leather, 1983; Miracle, 1989).

Tone 2 is produced with a low-to-high rising F0 and a relatively longer duration than Tone 1. Tone 3 is produced with a low-dipping F0 with a late raise, and a duration resembling that of Tone 2. For L2 learners and L1 children alike, Tone 2 and Tone 3 are the most difficult to master, partly due to these tones' more complex F0 contours (Leather, 1983; Li & Thompson, 1977; Shen & Lin, 1991; Sun, 1998; Wiener, 2017; Zhang, 2018). L2 learners in particular tend to produce Tone 2 and Tone 3 utterances that are judged as inaccurate by native listeners (Hao, 2012; Yang, 2015, 2016).

Tone 4 is produced with a high-falling F0 and typically the shortest mean duration. L2 learners tend to acquire Tone 4 earlier than Tone 2 and Tone 3 and generally produce Tone 4 with accuracy similar to that of Tone 1 (Hao, 2012; Yang, 2015; Zhang, 2018). When producing Tone 4, however, L2 learners often fail to sufficiently expand their pitch range to a native-like range (Shen, 1989).

Whereas the order of tone acquisition varies across L2 studies depending on the testing paradigm, stimuli, and participants' backgrounds (see Yang, 2015, for a review), Tone 1 and Tone 4 are generally acquired earlier and produced more accurately than Tone 2 and Tone 3, which L1 English–L2 Mandarin learners continue to struggle with even after multiple years of classroom learning (e.g., Hao, 2012). Given the challenges associated with L2 tone acquisition, there is increasing interest in exploring what potential pedagogical strategies may result in more native-like performance (Duff & Li, 2004). Most recently, researchers in this area have examined the use of adaptive training systems (Shih et al., 2010), web-based training platforms (Godfroid, Lin, & Ryu, 2017), visuospatial gestures (Morett & Chang, 2015), and incidental learning videogames (Wiener, Murphy, et al., 2019), among other approaches. We contribute to this ongoing discussion by investigating how explicit instruction of tone contours and high variability phonetic training affect the production of L2 Mandarin tones.

EXPLICIT INSTRUCTION OF TONE CONTOURS

In this study, we use the term *explicit instruction* to refer to instruction of tone that draws learners' attention to the F0 contours, which map to discrete phonological categories. We contrast this explicit instructional approach with what we call a nonexplicit approach in which learners must recognize the tonal contours and their phonological role through exposure to tonal minimal pairs and different learning tasks. Our approach expands on previous segmental explicit instruction studies that emphasized the place and manner of articulation for L2 segments, as well as similarities and differences between learners' L1 and L2 phonological systems (e.g., Arteaga, 2000; Kissling, 2013; Saito, 2011, 2013). This body of research has used a variety of production tasks (see Zampini, 2008, for a review) including naming, shadowing, and storytelling to demonstrate that by explicitly drawing learners' attention to specific acoustic features of the L2 sound system, learners' pronunciation accuracy—as judged by native speakers of the target language—typically improves relative to those who do not receive any instruction (e.g., DeKeyser, 2003; Jenkins, 2004; Kormos, 2014). Moreover, studies in this domain suggest that learning in a "nonexplicit" manner, as we define it, or implicitly (i.e., without awareness; see DeKeyser, 2003) forces learners to discover the relevant acoustic cues for themselves, which may delay L2 pronunciation improvement and potentially restrict learners from ultimately achieving native-like productions (e.g., Bongaerts et al., 1997; see also Lord, 2010, for a discussion of immersion and instruction).

Yet, despite the previously reported beneficial effects of explicit instruction on nonnative segmental production accuracy, prior evidence suggests that explicit instruction of suprasegmentals has a null effect on nonnative speakers' tone production accuracy. Chun, Jiang, and Ávila (2013) recorded first-year (third-quarter) L2 Mandarin classroom learners' ($N = 16$) tone productions before and after training on visual tonal contours similar to those plotted in Figure 1. The authors simultaneously presented learners with the visual F0 contours with audio input (both obtained from an L1 Mandarin speaker) and asked the learners to repeat the perceived input. The authors argued that by explicitly drawing the learners' attention to the intended F0 contour while listening to speech, learners could better approximate native speakers' F0 movement and range. Learners' mean production accuracy—as judged by L1 Mandarin listeners—showed a nonsignificant improvement from 83% (pretest) to 85% (posttest).

In a follow-up experiment involving a similar pre- and posttest recording design, Chun et al. (2015) examined productions of disyllabic words by students enrolled in an L2 Mandarin university course ($N = 35$, including nine non-English L1 speakers). Learners were given model productions from a native Mandarin speaker of the same gender and instructed to listen to the productions while viewing the visual tone contour with the software Praat (Boersma & Weenink, 2014). Learners were then instructed to record their own speech such that their utterance matched the perceived auditory tone and visual contour. Chun et al. once again predicted that by explicitly drawing learners' attention to the visual F0 contour, learners would approximate the target tone during production. As a result of explicit tone contour training, Tone 4 productions showed the most improvement while the other tones showed little to no improvement. Overall, the authors found no significant improvement between the pre- and posttest accuracy results, with mean production accuracy near 55%.

Whereas Chun et al.'s (2013, 2015) results suggest a null effect of instructional method, it is important to note that the tested participants were already familiar with Mandarin tones given their multiple quarters of L2 classroom instruction. Presumably all of the students tested had previously undergone some form of tone training in their classroom learning. Thus, these two studies did not test adults who were truly unfamiliar with the Mandarin tonal system, that is, the typical L2 learner in a beginner Mandarin classroom. Additionally, Chun et al. only exposed learners to productions from one speaker. The authors' results may have therefore been partly due to speaker-specific traits (e.g., a limited F0 range).

PHONETIC VARIABILITY IN SPEECH INPUT

Explicit phonetic instruction is often combined with perception practice involving input from native speakers. A corollary line of research has examined how the acoustic nature of such training input affects L2 learners' perception of the target speech sounds (e.g., Bradlow et al., 1997; Lively, Logan, & Pisoni, 1993; McCandliss et al., 2002). Studies in this domain have generally concluded that L2 learners benefit from exposure to acoustically varied speech—namely, high variability phonetic training (HVPT; see Thomson, 2012, for a review). Research using HVPT shows that

exposure to a wide range of cues typically improves learners' L2 perception of the target speech sound; learners must generalize over speaker-specific acoustic characteristics and extract common phonetic patterns that define category membership. In contrast, exposure to low variability input can limit learners' perception improvement because learners may struggle to separate speaker-specific idiosyncrasies from category-specific cues (Bradlow et al., 1999; Hardison, 2003; Thomson & Derwing, 2014; though see Jongman & Wade, 2007, for conflicting evidence).

L2 perception training can also affect L2 production, though our understanding of how these modalities interact remains unclear and is dependent upon numerous experimental and learner factors (see Sakai & Moorman, 2018, for a recent review). With respect to whether HVPT improves L2 learners' tone productions, limited evidence suggests L2 Mandarin learners' tone productions are facilitated by exposure to multi-talker speech. Wang and her colleagues (Wang, Jongman, & Sereno, 2003; Wang et al., 1999) recorded L1 English–L2 Mandarin learners ($N = 16$) with one to two semesters of classroom Mandarin experience reading aloud a list of Mandarin sounds written in Pinyin romanization. Learners then took part in a 2-week perceptual training program designed to improve their tone identification via exposure to four talkers. After the 2-week training was completed, participants read aloud the same list of Mandarin sounds. These pre- and posttraining productions were then rated by L1 Mandarin speakers. Learners' overall tone production accuracy improved by 18%. This improvement was observed across all four tones, though tone type differences emerged. Tone 1 productions were native-like in terms of F0 height and contour while Tone 4 productions improved but did not fully resemble native speakers' productions in terms of F0 height and fall. Although productions from Tone 2 and Tone 3 improved, these two tones improved the least, in part, due to these tones' more complex F0 contours.

In sum, Wang et al.'s findings suggest that exposing L2 learners to more varied tone contours may help learners generalize over speaker-specific characteristics and extract common phonetic patterns that define tone categories. Wang et al.'s findings, however, should be interpreted with caution as their participants, like the participants tested in Chun et al. (2013, 2015), were enrolled in an L2 Mandarin class at the time of testing and had at least one semester of classroom experience prior to the training session. The reported production improvement may have been modulated by their additional classroom input outside of the experiment and may not reflect the typical improvement seen in truly beginner L2 learners.

THE PRESENT STUDY

The present study advances previous L2 Mandarin speech production research in two ways. First, we acknowledge that evidence is dependent upon a list of learner and experimental conditions, including learners' individual aptitude, language background, experimental paradigms, training and testing materials, instructional techniques, use of feedback, and learning contexts, among other variables (e.g., Chang & Bowles, 2015; Faretta–Stutenberg & Morgan–Short, 2018; Hao & de Jong, 2016; Ke, 1998; Kingston, 2003; Kissling, 2013; Mok et al., 2018; Piske, MacKay, & Flege, 2001; Wiener, 2020). For example, Perrachione et al. (2011) demonstrated that for L1 English speakers learning to categorize L2 Mandarin tones, HVPT (i.e., speech from four talkers) facilitated tone learning but only for those participants with strong pitch perception abilities. Low variability phonetic training (i.e., speech from one talker) was more beneficial for participants with weak perceptual abilities (see also Sadakata & McQueen, 2014, for similar findings but Dong et al., 2019, for conflicting results). Given that we cannot control for all possible learning variables and individual differences, we take a "big data" approach to examine the putative benefits of explicit instruction and HVPT as they pertain to a wide range of adult nonnative Mandarin learners. We examined over 60,000 total productions from an L1 Mandarin group, an L2 Mandarin group, and a monolingual L1 English group. This allows us to observe general patterns that emerge over a diverse population of learners and put forth pedagogical recommendations that may impact the widest possible range of L2 learners.

Second, we used a Mandarin-like artificial language (see Ettlinger et al., 2016, for an artificial language review) that involves acquisition of sound–symbol mappings analogous to Chinese sound–character learning. Because purely auditory-only tone training runs the risk of task-specific strategies that may not reflect the processes and mechanisms involved in language learning (e.g., Chandrasekaran, Sampath, & Wong, 2010; Wong & Perrachione, 2007) and because orthographic information may aid L2 phonological and lexical learning (Escudero,

2015; Showalter & Hayes–Harb, 2013; though see Bassetti, 2006, 2007, for conflicting results), we trained participants on sound–symbol pairs that represent Mandarin-like syllable–tone words. More importantly, our artificial language allows us to compare performance across three groups with different prior tonal experiences (e.g., L1 experience, L2 classroom input, naïve listeners with no experience) while simultaneously controlling for exposure to our stimuli and lexical knowledge. We thus used an artificial language to simulate the L2 Mandarin acquisition process in a reduced period of time. This approach allows us to explore the following three research questions:

RQ1. Does explicit instruction of tone contours improve nonnative learners' Mandarin tone productions?
RQ2. Does high variability phonetic training improve nonnative learners' Mandarin tone productions?
RQ3. Do explicit instruction and high variability phonetic training interact to improve nonnative learners' Mandarin tone productions?

We tested L1 Mandarin, L2 Mandarin, and monolingual L1 English speakers using the speech shadowing task (see Bates & Liu, 1996). This simple task requires participants to repeat speech immediately after it is perceived, with participants mimicking the acoustic cues they deem most relevant. We recorded these utterances and asked L1 Mandarin listeners to categorize the productions into the four Mandarin tones. We expect that overall production accuracy—as judged by L1 Mandarin listeners—will reflect experience with Mandarin: The L1 Mandarin group will outperform the L2 Mandarin group who will in turn outperform the monolingual L1 English group (e.g., Hao, 2012; X. Wang, 2013; Wayland & Guion, 2004; Zhang, 2018). With respect to RQ1, L1 English participants explicitly trained may outperform those not explicitly trained given that explicit instruction may draw participants' attention to the F0 contour and its phonological role. If an effect is observed, this effect size may be relatively small given Chun et al.'s (2013, 2015) null findings. Productions of Tone 1 and Tone 4 may be produced more accurately than productions of Tone 2 and Tone 3 given that the former tones are typically acquired before the latter tones (e.g., Yang, 2015) and that L1 English speakers are initially more sensitive to tone height than tone contour (e.g., Huang & Johnson, 2010). Learners explicitly trained on tone contours may also show greater accuracy on Tone 2 and Tone 3 productions than learners not explicitly trained given that these perceptually similar tones may benefit the most from visual cues (e.g., Mok et al., 2018).

With respect to RQ2, L1 Mandarin and L2 Mandarin groups should show no sensitivity to talker variability because these participants will already be familiar with the tone categories at the time of testing. In contrast, L1 English participants exposed to multiple talkers may show greater accuracy than those exposed to a single talker given that high variability input typically leads to improved productions (e.g., Wang et al., 2003). If an effect is observed, learners exposed to greater talker variability may produce tones more accurately overall, or only specific tones, such as the earlier acquired Tone 1 and Tone 4.

With respect to RQ3, an interaction may occur such that high variability input is only beneficial if combined with explicit instruction of tone categories. Participants made explicitly aware of the tone category and then presented with varied input, may better learn both the primary F0 cue and potential secondary acoustic cues (e.g., duration, amplitude) than those participants presented with varied input but not made explicitly aware of the contours. This explicit awareness combined with varied speech exemplars may be especially helpful when learning to produce the more challenging Tone 2 and Tone 3 categories, which share overlapping F0 contours and therefore benefit from accurate secondary cues such as F0 turning point. In contrast, listeners who are both unaware of the specific F0 contours for each category and presented with varied input may struggle to identify the relevant acoustic cues that define the four tone categories.

METHOD

Participants

Forty native Mandarin speakers from central and northern mainland China served as the L1 Mandarin group. All L1 Mandarin speakers self-reported speaking only Mandarin and no other regional dialect. All L1 Mandarin speakers were living in the United States at the time of testing and were proficient L2 English speakers. Forty native English speakers learning Mandarin as an L2 in a university setting served as the L2 Mandarin group. All L2 learners had completed an estimated 140 classroom hours and 140 self-study

hours prior to the experiment and were enrolled in an intermediate Mandarin course at the time of testing. Eighty self-identified monolingual English speakers with no prior experience studying Mandarin or any tonal language served as the L1 English beginner group. An additional 104 L1 Mandarin–L2 English speakers from mainland China served as tone raters. All raters self-reported speaking only Mandarin (i.e., no other regional dialect) and did not participate in the speaking tasks.

All 264 participants and raters were undergraduate or graduate students at a public university in the United States and had normal hearing and vision. Participants' mean age was 21.4 (range: 18–35). All participants and raters were paid $10 per hour.

Stimuli

To control the exposure to the stimuli across all three groups, an artificial Mandarin-like tonal language was designed. The language consisted of 24 monosyllabic consonant–vowel (CV) syllables that did not violate Mandarin phonotactics but which are absent in modern Mandarin and thus processed as nonwords (e.g., *fe*, which is analogous to the English phonotactically legal nonce word "blick"; see H. S. Wang, 1998; Wiener & Turnbull, 2016). Each syllable was produced with one or more of the four Mandarin tones (Figure 1) by four phonetically trained native Mandarin talkers (two male; two female). CV syllables varied in token frequency, co-occurrence with tone type, and number of tonal homophones. Unique to our experimental design, the stimuli replicated the natural syllable–tone asymmetry, syllable neighborhood density, and syllable–tone homophone density that L2 learners (and L1 speakers) are exposed to in spoken Mandarin (see DeFrancis, 1984; Duanmu, 2007, 2008, and Wiener & Ito, 2015, 2016, for additional information). This included homophonous items that were differentiated only by their visual form (analogous to English /tu/ as "two," "to," and "too"). Because this study is part of a larger series of studies involving statistical learning, we refer the reader to Wiener (2015) and Wiener, Ito, and Speer (2016, 2018) for additional information including acoustic measurements and the full stimuli. In total, 130 syllable–tone "words" were created with each syllable–tone paired with a unique black and white nonce symbol (see Figure 3). These nonce symbols were designed to simulate the challenge of sound-to-Chinese-character acquisition and force participants to form word-like representations in which the tone's phonological role was necessary for symbol identity.

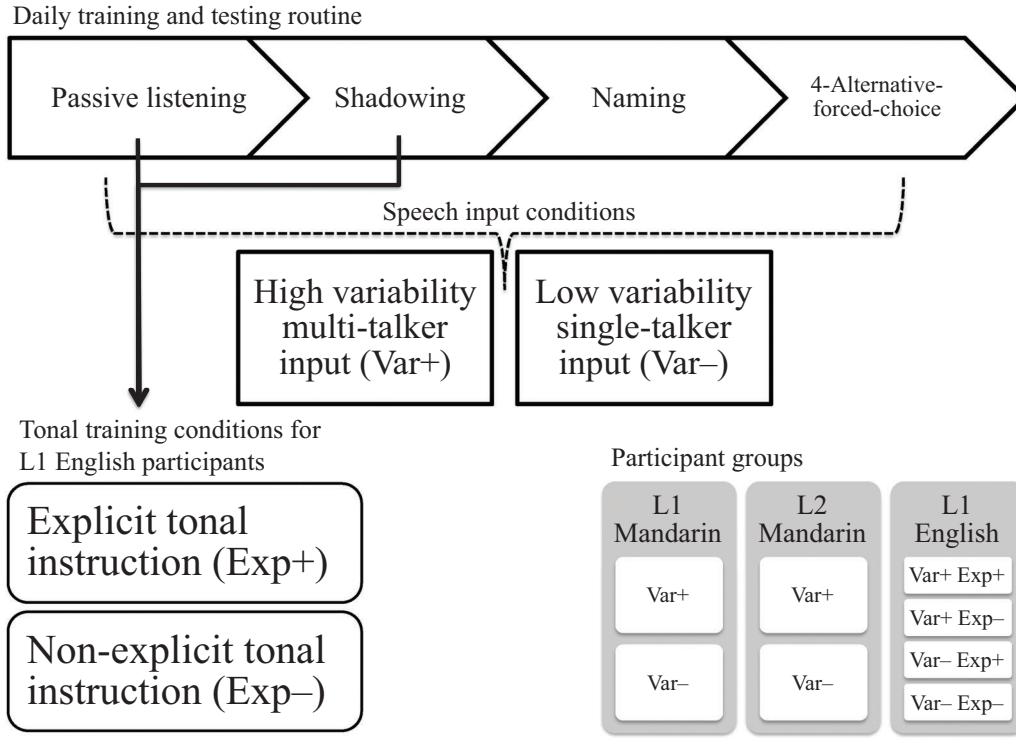
Procedure

Participants followed the same 30-minute training and testing routine for 4 consecutive days (see Figure 2, top). First, participants were seated in a sound-attenuated booth and presented with 131 trials of self-paced passive listening. On each trial, the target word's audio was presented over headphones while the nonce symbol was simultaneously displayed on a monitor. Participants were instructed to remember the sound–symbol pair. Next, participants were presented with 131 new trials of self-paced shadowing. This shadowing task, which is the focus of this article, presented the nonce symbol on screen while its audio label was simultaneously presented over headphones. After the audio was played, participants were instructed to repeat the perceived word as clearly and accurately as possible while the nonce symbol remained on the screen. After producing the word, participants clicked the mouse to continue to the next trial. All recordings were made using Praat (Boersma & Weenink, 2014) at 44.1 kHz with 16-bit resolution. In total, each participant produced 524 tokens in the shadowing task across 4 consecutive days. After shadowing, participants performed a word-naming (with feedback) task, in which a nonce symbol was presented on screen and participants had to produce its syllable–tone label. The last daily experimental task was forced choice word identification with eye-tracking (with feedback), in which four symbols were presented on screen while a syllable–tone target was presented over headphones. In this last task, participants were asked to mouse-click on the symbol matching the perceived audio while their eye movements were recorded (see Wiener et al., 2018, for additional details).

To manipulate the phonetic input, half of the participants in each group (including the L1 Mandarin and L2 Mandarin groups) shadowed all four talkers in the high variability training condition (Var+) while the other half of the participants only shadowed one female talker in the low variability training condition (Var–).

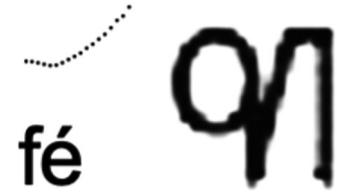
Because the L1 Mandarin and L2 Mandarin groups were already familiar with Mandarin tones, only the monolingual L1 English group took part in the explicit instructional method manipulation. Half of the L1 English participants began each day of training (prior to the passive listening task) with a 5-minute computerized, self-paced, explicit lesson on the four tones and their

FIGURE 2
Experimental Design



F0 height and contour characteristics (Exp+). This instruction used the Mandarin *ma* quadruplet to introduce the four tones and their visual F0 information following the typical pedagogical approach used in L2 classrooms and textbooks when introducing monosyllables in isolation (see Shen, 1989; Xing, 2006; Yang, 2015). Because the tone number was irrelevant to learning the artificial language, the terms *high-level tone*, *rising tone*, *low-dipping tone*, and *falling tone* were used to explain the four tones. Participants were instructed to pay attention to the pitch or tone of the syllable and then simultaneously presented with an auditory syllable–tone over headphones while its Pinyin was displayed onscreen with tone diacritics. After participants mouse-clicked on the symbol, the auditory stimulus was presented again while the syllable's rendered F0 contour was simultaneously displayed on the computer monitor (generated through Praat). At the end of the computerized lesson, participants were again reminded that the pitch or tone of the syllable is important for symbol learning. During the passive listening and shadowing phases of the training, the word's visual tone contour (following Liu et al., 2011), Pinyin with tone diacritics (following Mok et al., 2018; Showalter & Hayes–Harb, 2013), and nonce symbol were all simultaneously shown on screen while the audio was presented over headphones. Figure 3 shows an example of *fe2* with its nonce symbol on the right, the tone's rising F0 contour, and the word's syllable–tone written in Pinyin romanization on the left.

FIGURE 3
Explicit Instruction Slide for *fe2* Containing Pinyin and Tone Contour (Left) and Nonce Symbol (Right)



The other half of the L1 English participants were assigned to the nonexplicit training condition (Exp–). In this condition, no tone training was provided before the tasks and only the nonce symbol was displayed on screen with the audio during the passive listening and shadowing tasks

(i.e., only the nonce symbol on the right-hand side of Figure 3 was shown; the Pinyin and tone visualization were not shown).

Participants' shadowing recordings were first cleaned and trimmed. All problematic files in which participants did not speak clearly or follow the instructions were removed from the dataset. The remaining 61,848 tokens were presented to the 104 L1 Mandarin raters (none of whom participated in the language learning task). Raters were told they would hear Mandarin-like speech and to categorize each utterance based on its perceived tone by pressing 1, 2, 3, or 4 on a keyboard. All utterances were presented to raters over headphones in a sound-attenuated booth. Each rater heard 600 pseudo-randomized trials across two blocks with a 2-second interstimulus interval. These trials presented roughly the same syllable–tone distribution as that presented to the participants during training and testing. Raters heard at least three utterances from all 160 participants. There was moderate agreement among the raters, $K = .74$, 95% confidence interval (CI) [.73, .75]. Our measure of production accuracy therefore represents whether L1 Mandarin listeners perceived the utterance as the intended tone or a different tone.

RESULTS

Between-Group Analyses

Group production accuracy across all 4 days was first calculated. Ninety-five percent CIs revealed that the L1 Mandarin speakers were most accurate [.86, .89]. L2 Mandarin learners were second-most accurate [.77, .80], while L1 English monolinguals were least accurate [.69, .74]. For all three groups, high and low variability conditions resulted in relatively overlapping 95% CIs: L1 Mandarin Var+ [.85, .89], Var– [.86, .90]; L2 Mandarin Var+ [.76, .80], Var– [.76, .82]; L1 English Var+ [.68, .74], Var– [.68, .76]. To test whether group, day of training, speaker variability, and their interaction resulted in statistically different log odds of correct identification (where each response was coded as 1 if correct and as 0 if incorrect), a mixed-effects logistic regression model was built using the lme4 package (Bates et al., 2014) in R (version 3.3.3; R Core Team, 2017). The most parsimonious model was determined using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017), which tested whether the inclusion of variables and their interaction improved the model fit. Speaker variability was contrast coded, day was treated as a continuous variable, and group was dummy coded with the L2 Mandarin group as the reference level—this allowed for two planned contrasts involving the L2 Mandarin group to the L1 Mandarin group and the L2 Mandarin group to the monolingual L1 English group. Reported estimates reflect effect sizes and direction of effects. Model formula:

$$\begin{aligned}&\text{glmer}(\text{accuracy} \sim \text{group} \times \text{variability} \times \text{day}\\&\quad + [\text{group}|\text{rater}] + [\text{day}|\text{item}],\\&\quad \text{family} = \text{“binomial”})\end{aligned}$$

A main effect of group was found, confirming that the L1 Mandarin group's productions were identified more accurately than the L2 group's productions ($\beta = 0.756$, $SE = 0.12$, $Z = 6.25$, $p < .001$), and that the L2 Mandarin group's productions were identified more accurately than the L1 English group's productions, $\beta = -0.406$, $SE = 0.10$, $Z = -4.02$, $p < .001$. Neither a main effect of speaker variability, $\beta = -0.005$, $SE = 0.19$, $Z = -0.03$, $p = .97$, nor its interaction with group or day was found (p's $> .05$). A null effect of day of training was found, $\beta = 0.017$, $SE = 0.01$, $Z = 1.68$, $p = .10$. No interaction with day was found (p's $> .05$).

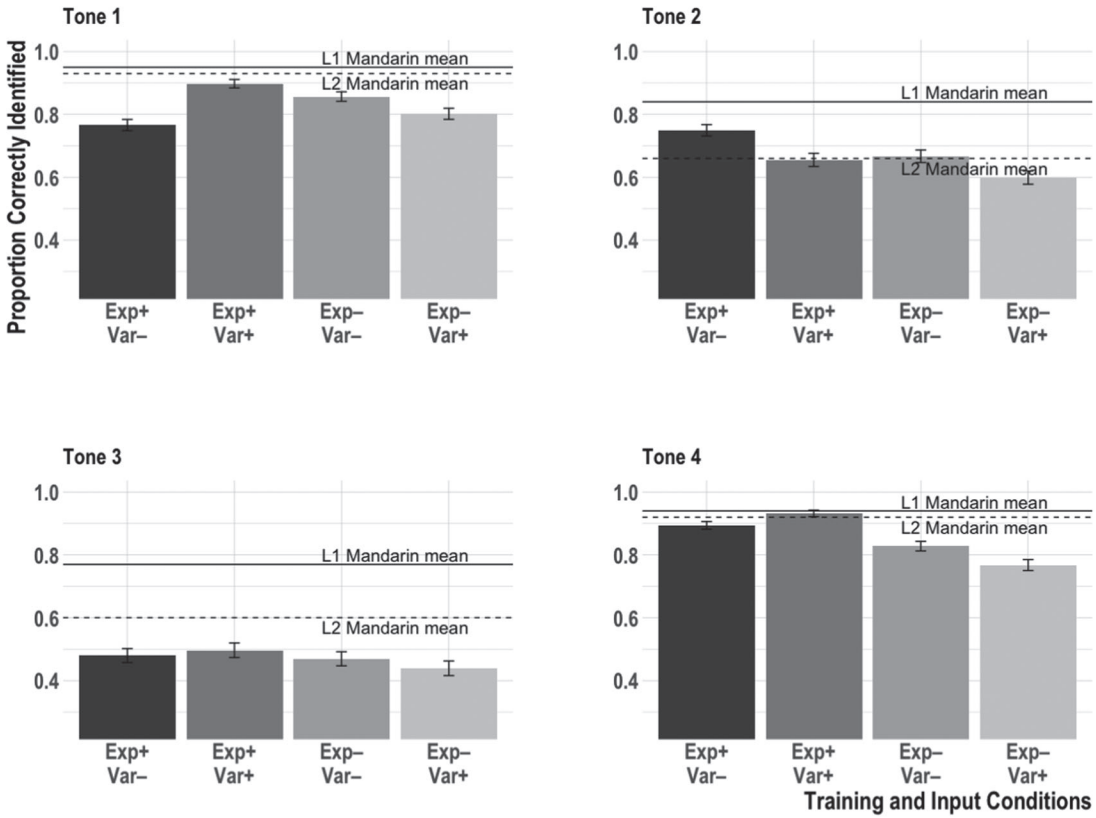
To confirm that the group differences were robust over tone type, subset analyses of each tone were carried out. The L1 Mandarin group's productions of Tone 2 and Tone 3 were more accurate than those of the L2 Mandarin group (p's $< .05$), but the two groups produced Tone 1 ($p = .85$) and Tone 4 ($p = .89$) with similar accuracies. The L2 Mandarin group's productions of Tone 1, Tone 3, and Tone 4 were more accurate than the L1 English group's productions (p's $< .05$), but the two groups produced Tone 2 with similar accuracies ($p = .90$).

Figure 4 plots the L1 English participants' results by tone type across the four conditions (with error bars indicating 95% CIs) along with means for the L1 Mandarin (solid line) and L2 Mandarin (dotted line) groups. Figure 4 highlights the between-group findings: the L2 Mandarin group achieved native-like Tone 1 and Tone 4 productions, but not native-like Tone 2 and Tone 3 productions. The L2 Mandarin group outperformed the L1 English group in Tone 1 and Tone 3 productions, partially in Tone 4 productions, but not in Tone 2 productions.

L1 English Group Analyses

To further explore the monolingual L1 English data and test the effect of explicit instruction, a

FIGURE 4
L1 English Speakers' Mean Tone Production Accuracy by Tone and Training Condition With L1 Mandarin and L2 Mandarin Group Means



Note. Var+ = high variability training condition; Var– = low variability training condition; Exp+ = explicit training condition; Exp– = nonexplicit training condition.

new model was built testing whether accuracy differed across the four tones (dummy coded with Tone 1 as the reference level). Speaker variability, instructional method, and their interaction (both variables contrast coded) were included as fixed effects:

glmer(accuracy ~ tone + variability
× instruction + [variability × instruction|rater]
+ [tone|item], family = "binomial")

A main effect of instruction was found; productions from participants explicitly trained were more accurate than productions from participants not explicitly trained, $\beta = 0.318$, $SE = 0.13$, $Z = 2.40$, $p = .02$. A null effect of speaker variability was found, $\beta = -0.053$, $SE = 0.06$, $Z = -0.81$, $p = .41$. A two-way interaction between instructional method and speaker variability was found, $\beta = -0.363$, $SE = 0.16$, $Z = -2.27$, $p = .02$. A main effect of tone was also found: All tone pairwise comparisons were significantly different from one another (p's < .01). Ninety-five percent CIs revealed that Tone 4 was most accurate [.85, .86], followed by Tone 1 [.82, .83], followed by Tone 2 [.66, .70], with Tone 3 least accurate [.46, .48].

To identify the locus of the two-way interaction and confirm that the main effect of instructional method was not driven by productions of a particular tone, a series of post hoc models were built for each tone type. For each model, variability and instruction method were contrast coded. Model formula:

glmer(accuracy ~ variability × instruction
+ [variability × instruction|rater] + [1|item],
family = "binomial")

For Tone 1, no main effects were found: Learners' productions from both explicit/nonexplicit conditions did not differ in accuracy, $\beta = 0.068$, $SE = 0.27$, $Z = 0.25$, $p = .80$. Similarly, learners' productions from both high and low variability input conditions did not differ in accuracy, $\beta = -0.183$, $SE = 0.19$, $Z = -0.95$, $p = .34$. A two-way interaction between variability and instruction was found, $\beta = 0.681$, $SE = 0.27$, $Z = 2.48$, $p = .01$. Productions from learners given explicit instruction and high variability input (Exp+ Var+) were more accurate than productions from learners given the same explicit instruction but low variability input (Exp+ Var–): $\beta = 1.006$, $SE = 0.42$, $Z = 2.34$, $p = .02$. In contrast, no accuracy difference was found for the nonexplicit instruction conditions regardless of the variability in the input (p's > .2).

For Tone 2, a main effect of instruction was found: Productions from learners given Exp+ instruction were more accurate than productions from learners given Exp– instruction, $\beta = 0.343$, $SE = 0.14$, $Z = 2.36$, $p = .02$. A null effect of variability, $\beta = -0.181$, $SE = 0.10$, $Z = -1.77$, $p = .08$, and a null interaction, $\beta = -0.087$, $SE = 0.14$, $Z = -0.60$, $p = .55$, were found.

For Tone 3, null effects of instruction, $\beta = 0.137$, $SE = 0.14$, $Z = 0.97$, $p = .32$; variability, $\beta = -0.027$, $SE = 0.09$, $Z = -0.35$, $p = .78$; and their interaction, $\beta = 0.054$, $SE = 0.14$, $Z = 0.38$, $p = .70$, were found. Productions from all four conditions were equally accurate.

For Tone 4, a main effect of instruction was found: Productions from learners given Exp+ instruction were more accurate than productions from learners given Exp– instruction, $\beta = 1.022$, $SE = 0.29$, $Z = 3.50$, $p < .001$. Neither a main effect of variability, $\beta = -0.314$, $SE = 0.20$, $Z = -1.54$, $p = .12$, nor its interaction with instructional method, $\beta = 0.432$, $SE = 0.29$, $Z = 1.48$, $p = .14$, was found.

To summarize, a main effect of instructional method was found suggesting an overall accuracy improvement for those explicitly trained (RQ1). Subset analyses, however, revealed that that effect was primarily driven by Tone 2 and Tone 4 productions. For Tone 1 and Tone 3 productions, explicit instruction alone had a null effect on accuracy. Overall, input variability had a null effect as productions from learners trained on multiple talkers were as likely to be correctly identified as productions from learners trained on a single talker (RQ2). Finally, a two-way interaction was found (RQ3): for Tone 1 productions, high variability input was beneficial only when combined with explicit instruction. For Tone 2, Tone 3, and Tone 4 productions, no such two-way interaction was found.

DISCUSSION

Producing native-like speech as an adult L2 learner is not an easy task. Although lab-based research into adult L2 acquisition has made enormous strides in identifying the specific problems learners face, methodological improvement and collection of larger data sets are crucial for further advancement of pedagogies for L2 sound acquisition (King & Mackey, 2016; Marsden et al., 2018). In this study we used an artificial, Mandarin-like tonal language to test a large group of truly beginner nonnative adults unfamiliar with Mandarin

tone. We examined how explicit instruction of Mandarin tone contours, HVPT, and these variables' potential interaction affect the nonnative production of Mandarin tones. We present three findings from this study.

First, we found evidence that explicit instruction of tone facilitates L2 pronunciation improvement for Tone 2 and Tone 4 productions. By becoming explicitly aware of the F0 rise (Tone 2) and fall (Tone 4) involved with these tones, learners produced the tones more accurately than those learners who were not made explicitly aware of the F0 contours. To our knowledge, this is the first large-scale study to demonstrate a significant facilitatory effect of explicit instruction on truly beginner, nonnative adult learners' tone productions. These findings are in line with previous explicit instruction studies (e.g., DeKeyser, 2003; Hulstijn, 2005; Saito, 2011, 2013) and strengthen the claim that explicit instruction of difficult-to-acquire L2 cues can improve learners' speech productions. We note, however, that explicit instruction alone was not effective in improving learners' productions of Tone 1 or Tone 3.

Second, we did not find evidence that high variability input alone facilitates learners' tone productions. Productions from those trained on more variable input were not statistically more accurate than productions from those trained on less variable input. For the L1 Mandarin and L2 Mandarin groups, this finding was expected given their prior experience with the four tones. For the monolingual L1 English group, however, this was an unexpected result given the robust improvement observed in Wang et al. (2003), and the similar facilitatory effects reported in the lexical tone perceptual learning literature (e.g., Barcroft & Sommers, 2005, 2014; Sadakata & McQueen, 2014; Shih & Lu, 2015). The difference between the present study and Wang et al. may reflect the more complicated 4-day training paradigm involving nonce symbols paired with sounds, which presumably added a cognitive load that was absent in Wang et al.'s study.

Additionally, we tested truly beginner adults who were not currently involved in L2 classroom learning. The facilitatory effect of speaker variability in Wang et al. (2003) may have partially reflected the additional input participants received outside of the experiment during their L2 classroom learning. The difference between the two studies highlights the value in using a variety of stimuli and tasks in assessing native and nonnative speech (e.g., Marsden et al., 2018; Perrachione et al., 2011).

We note that although the predicted accuracy improvement was not observed, high variability input did not result in overall statistically *less* accurate productions; HVPT was certainly not detrimental to participants' learning. Moreover, HVPT interacted with explicit instruction to facilitate Tone 1 productions. For Tone 1 productions, exposure to varied input alone or explicit instruction alone was not sufficient for Tone 1 improvement. Rather, it was the interaction of the two variables: Becoming explicitly aware of F0 height cues and being exposed to varied, multi-speaker exemplars resulted in more accurate Tone 1 productions. This finding is in line with Shih and Lu's (2015) claim that learners can sometimes struggle to recognize Tone 1 when the utterance is not entirely level. One interpretation of these results is that the term level may be misleading as many Tone 1 productions contain a slight F0 rise or fall (Ho, 1976; Howie, 1976; Tseng, 1981). For instance, our female talker used in the Var– condition demonstrated an average F0 increase of 20.8 Hz across all her utterances (as measured by F0 onset and offset of the vowel using Praat). Without HVPT input to show just how "level" Tone 1 is, learners may have problems accurately perceiving and producing the correct tone.

Finally, our results support previous L2 phonetic studies that demonstrated Tone 1 and Tone 4 are generally the first tones acquired by L2 learners, while Tone 2 and Tone 3 tend to be more difficult and require additional time and input to master (e.g., Hao, 2012; Wang et al., 1999). For both L2 Mandarin and L1 English speakers alike, Tone 3 remained the most difficult tone to accurately produce, though the L2 Mandarin group was more accurate than the monolingual L1 English group, presumably due to their L2 classroom experience with Tone 3. In contrast, while both nonnative groups produced Tone 2 less accurately than the L1 Mandarin group, no overall difference between the L2 Mandarin and L1 English groups was found. These results suggest Tone 2 productions may be more malleable to immediate improvement, given the right explicit training conditions. Yet, despite the observed Tone 2 improvement, our results suggest a lengthy learning plateau may follow. The L2 Mandarin participants appeared to plateau in their Tone 2 (and Tone 3) production improvement despite over a year of previous classroom exposure.

Unfortunately for nonnative learners, our relatively poor Tone 2 and Tone 3 results are in line with previous lab-based and classroom-based L2 Mandarin acquisition research (e.g., Chen et al., 2016; Everson & Shen, 2010; Hao, 2012; Wiener,

2017; Wiener, Lee, & Tao, 2019; Yang, 2015, 2016). Even the L2 Mandarin speakers tested in the present study with over 140 hours of classroom experience still produced Tone 2 and Tone 3 with less than 70% accuracy. Yet, we note that many of our L1 Mandarin speakers also produced Tone 2 and Tone 3 utterances that were incorrectly judged by L1 Mandarin raters (see Shen & Lin, 1991, for corroborating evidence). The reality is these two tones are challenging for all speakers—even L1 children do not typically demonstrate tone mastery until after the age of four or five (Singh & Fu, 2016; Wong, 2013). It may be misguided for educators and researchers to expect L2 tone mastery after a semester or two of limited classroom practice. There is a real need for classroom-based longitudinal research to better understand the acquisition process of adult L2 tone productions. More importantly, such research may help educators and researchers set more evidence-based expectations and goals for L2 learners.

Taken together, our findings motivate the following recommendation for L2 Mandarin tone pronunciation instruction. First, learners should continue to be explicitly instructed on F0 contours and the lexical role of tone as current pedagogy suggests (e.g., Xing, 2006; Yang, 2015). Our results show that for L1 English speakers, explicit instruction led to proportionally more accurate productions across all four tones. We found fairly robust effect sizes for Tone 2 and Tone 4, and for Tone 1 when combined with high phonetic variability input.

Similarly, whereas the facilitatory nature of high variability input was not observed in the present study, we see no reason not to expose learners to highly varied tone productions from multiple speakers. However, we note that such exposure should be delayed and gradually introduced only after learners are explicitly aware of tone's F0 contours and phonological role. When high variability input was combined with explicit instruction, participants produced the proportionally most accurate productions for Tone 1, Tone 3, and Tone 4. In contrast, when high variability input was combined with non-explicit instruction (i.e., learners were unaware of what acoustic cues to attend to and the cues were highly variable), participants produced the proportionally least accurate production for Tone 2, Tone 3, and Tone 4. An explicit instruction approach combined with an adaptive and engaging training paradigm (e.g., Shih et al., 2010; Wiener, Murphy, et al., 2019) could therefore initially deliver low variability, single-speaker input followed by a gradual introduction of high variability, multispeaker speech involving Tone 1 and Tone 4 exemplars before introducing Tone 2 and Tone 3 exemplars. Future research may explore how steadily introducing more variable phonetic input over time affects L2 learners' productions—including productions in running speech (see Tseng, 1981; Tseng et al., 2005)—and to what degree tone productions improve given multisyllabic training (Chang & Bowles, 2015).

Although we took a "big data" approach to understanding L2 production of tone, this study is not without limitations. First, we acknowledge that individual differences in learners' abilities may have contributed to our results. Musical experience, pitch perception, phonological awareness, working memory, and numerous other learner-specific traits may have partly driven the observed patterns (e.g., Bowles, Chang, & Karuzis, 2016; Faretta–Stutenberg & Morgan–Short, 2018; Perrachione et al., 2011). Future studies will need to control for individual differences with greater rigor.

Second, whereas we found that explicit instruction of tone resulted in more accurate Tone 2 and Tone 4 productions, we cannot say for certain whether this improvement was due to the short daily tone lesson, the visual F0 contours, the inclusion of the Pinyin romanization, or some combination of the factors. Future studies will need to explore the relative contribution of each explicit learning aid and examine whether other explicit approaches, including novel orthography (e.g., Hayes–Harb & Cheng, 2016) yield even greater improvement.

Third, our results from the low variability condition may have partly been driven by speaker-specific idiosyncrasies. Thus, like Chun et al.'s (2013, 2015) study, our study cannot determine how much learning (or lack thereof) was driven by the low phonetic variability in the stimuli versus input from only one female talker's productions. Future work will need to carefully tease apart these two accounts.

Finally, our raters were exposed to an incredibly high degree of variability across 160 participants. On the one hand, one goal of L2 acquisition is to be accurately understood by a first-time listener. Our novel approach involving fewer exemplars per speaker may better model the natural intelligibility of L2 speech by L1 listeners. On the other hand, our approach clearly affected the overall rater accuracy, (cf. Chang, Yao, & Huang, 2017) and contributed to only moderate agreement among raters. We recognize that if raters were given more exemplars from each

participant, and thus had become more familiar with the acoustics of each speaker's voice, we may have found a different pattern of results.

In conclusion, the present study demonstrated that explicit instruction of tonal contours improves L1 English speakers' productions of the rising and falling Mandarin tone contours. Although high variability training alone did not directly improve participants' productions, a general trend was found in which explicit instruction combined with high variability input resulted in proportionally more accurate productions. High variability input combined with explicit instruction can be particularly effective for improving the productions of high-level tone contours. These results motivate the recommendation of an explicit approach to tone teaching combined with gradual exposure of more variable, multi-speaker speech delivered through an adaptive training program.

ACKNOWLEDGMENTS

This research was supported by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (BCS-1451677) to the first and third authors. We are incredibly grateful to James Brennan, Raphael Fan, and Xuyu Li for their help with the experiments; to all the participants who took part in this study; and to the thoughtful *Modern Language Journal* reviewers for their substantial comments on earlier versions of this work.

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://osf.io/pxd5f/.

REFERENCES

ACTFL. (2012). *2012 annual research report*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

Arteaga, D. (2000). Articulatory phonetics in the first-year Spanish classroom. *Modern Language Journal, 84*, 339–354.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*, 387–414.

Barcroft, J., & Sommers, M. S. (2014). Effects of variability in fundamental frequency on L2 vocabulary learning: A comparison between learners who do and do not speak a tone language. *Studies in Second Language Acquisition, 36*, 423–449.

Bassetti, B. (2006). Orthographic input and phonological representations in learners of Chinese as a foreign language. *Written Language & Literacy, 9*, 95–114.

Bassetti, B. (2007). Effects of hanyu pinyin on the pronunciation of learners of Chinese as a foreign language. In A. Guder, J. Xin, & Y. Wan (Eds.), *The cognition, learning and teaching of Chinese characters* (pp. 156–179). Beijing, China: Beijing Language and Culture University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. Accessed 1 February 2015 at https://arxiv.org/abs/1406.5823

Bates, E., & Liu, H. (1996). Cued shadowing. *Language and Cognitive Processes, 11*, 577–582.

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics, 18*, 37–49.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (Version 5.4) [Computer program]. Accessed 1 October 2014 at http://www.praat.org/

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19*, 447–465.

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning, 66*, 774–808.

Bradlow, A. R., Akahane–Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*, 977–985.

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

Cai, J., Chen, J., & Wang, C. (Eds.). (2010). *Teaching and learning Chinese: Issues and perspectives*. Charlotte, NC: Information Age Publishing.

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

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America, 138*, 3703–3716.

Chang, Y. H. S., Yao, Y., & Huang, B. H. (2017). Effects of linguistic experience on the perception of high-variability nonnative tones. *The Journal of the Acoustical Society of America*, *141*, EL120–EL126.

Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of nonnative Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. *Speech Communication*, *84*, 46–56.

Chun, D. M., Jiang, Y., & Ávila, N. (2013). Visualization of tone for learning Mandarin Chinese. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th pronunciation in second language learning and teaching conference* (pp. 77–89). Ames, IA: Iowa State University.

Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation*, *1*, 86–114.

DeFrancis, J. (1984). *The Chinese language: Fact and fantasy*. Honolulu, HI: University of Hawai'i Press.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 313–348). Oxford, UK: Blackwell.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–397.

Dong, H., Clayards M., Brown H., & Wonnacott E. (2019). *The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones*. Accessed 14 October 2019 at https://doi.org/10.7717/peerj.7191

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). New York: Oxford University Press.

Duanmu, S. (2008). *Syllable structure: The limits of variation*. New York: Oxford University Press.

Duff, P. A., & Li, D. (2004). Issues in Mandarin language instruction: Theory, research, and practice. *System*, *32*, 443–456.

Escudero, P. (2015). Orthography plays a limited role when learning the phonological forms of new words: The case of Spanish and English learners of novel Dutch words. *Applied Psycholinguistics*, *36*, 7–22.

Ettlinger, M., Morgan–Short, K., Faretta–Stutenberg, M., & Wong, P. C. (2016). The relationship between artificial and second language learning. *Cognitive Science*, *40*, 822–847.

Everson, M., & Shen, H. (2010). *Research among learners of Chinese as a foreign language*. Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i at Monoa.

Faretta–Stutenberg, M., & Morgan–Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second Language Research*, *34*, 67–101.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995) Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*, 3125–3134.

Gandour, J. T. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, *11*, 149–176.

Godfroid, A., Lin, C. H., & Ryu, C. (2017). Hearing and seeing tone through color: An efficacy study of web-based, multimodal Chinese tone perception training. *Language Learning*, *67*, 819–857.

Goldberg, D., Looney, D., & Lusin, N. (2015). *Enrollments in languages other than English in United States institutions of higher education, Fall 2013*. Accessed 24 January 2020 at https://www.mla.org/content/download/31180/1452509/EMB_enrllmnts_nonEngl_2013.pdf

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*, 269–279.

Hao, Y. C. (2018a). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication*, *97*, 32–42.

Hao, Y. C. (2018b). Second language perception of Mandarin vowels and tones. *Language and Speech*, *61*, 135–152.

Hao, Y. C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*, 495–522.

Hayes–Harb, R., & Cheng, H. W. (2016). The influence of the Pinyin and Zhuyin writing systems on the acquisition of Mandarin word forms by native English speakers. *Frontiers in Psychology*, *7*, 785.

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, *33*, 353–367.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. New York/Cambridge: Cambridge University Press.

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

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition*, *27*, 129–140.

Jenkins, J. (2004). Research in teaching pronunciation and intonation. *Annual Review of Applied Linguistics*, *24*, 109–125.

Jongman, A., & Wade, T. (2007). Acoustic variability and perceptual learning: The case of non-native accented speech. In O.–S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning* (pp. 135–150). Amsterdam: John Benjamins.

Ke, C. (1998). Effects of language background on the learning of Chinese characters among foreign language students. *Foreign Language Annals*, *31*, 91–102.

Ke, C. (2018). *The Routledge handbook of Chinese second language acquisition*. New York: Routledge.

King, K. A., & Mackey, A. (2016). Research methodology in second language studies: Trends, concerns, and new directions. *Modern Language Journal*, *100*, 209–227.

Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, *46*, 295–348.

Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *Modern Language Journal*, *97*, 720–744.

Kormos, J. (2014). *Speech production and second language acquisition*. New York: Routledge.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Laver, J. (1994). *Principles of phonetics*. New York/Cambridge: Cambridge University Press.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, *11*, 373–382.

Lee, J., Jang, J., & Plonsky, L. (2014). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366.

Li, C. N., & Thompson, S. A. (1977). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language*, *4*, 185–199.

Li, Y., Wen, X., & Xie, T. (2014). CLTA 2012 survey of college-level Chinese language programs in North America. *Journal of the Chinese Language Teachers Association*, *49*, 1–49.

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning*, *61*, 1119–1141.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*, 1242–1255.

Lord, G. (2010). The combined effects of immersion and instruction on second language pronunciation. *Foreign Language Annals*, *43*, 488–503.

Marsden, E., Morgan–Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*, 321–391.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 89–108.

Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association*, *24*, 49–65.

Mok, P. P. K., Lee, A., Li, J. J., & Xu, R. B. (2018). Orthographic effects on the perception and production of L2 mandarin tones. *Speech Communication*, *101*, 1–10.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, *102*, 1864–1877.

Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*, 347–353.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97.

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

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*, *54*, 51–78.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*, 191–215.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*, 1318.

Saito, K. (2011). Examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: An instructed SLA approach to L2 phonology. *Language Awareness*, *20*, 45–59.

Saito, K. (2013). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition*, *35*, 1–29.

Saito, K., & Wu, X. (2014). Communicative focus on form and second language suprasegmental learning: Teaching Cantonese learners to perceive Mandarin tones. *Studies in Second Language Acquisition*, *36*, 647–680.

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

Shen, X. S. (1989). Toward a register approach in teaching Mandarin tones. *Journal of Chinese Language Teachers Association*, *24*, 27–47.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, *34*, 145–156.

Shih, C., & Lu, H. Y. D. (2015). Effects of talker-to-listener distance on tone. *Journal of Phonetics*, *51*, 6–35.

Shih, C., Lu, H. Y. D., Sun, L., Huang, J. T., & Packard, J. (2010). An adaptive training program for tone acquisition. In *Proceedings of the 5*th *International Conference on Speech Prosody* (paper 981). Baixas, France: International Speech Communication Association.

Singh, L., & Fu, C. S. (2016). A new view of language development: The acquisition of lexical tone. *Child Development*, *87*, 834–854

Showalter, C. E., & Hayes–Harb, R. (2013). Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research*, *29*, 185–200.

Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese*. Honolulu, HI: University of Hawai'i Press.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer mediated approach. *Language Learning*, *62*, 1231–1258.

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*, 326–344.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1–30.

Tseng, C. Y. (1981). *An acoustic phonetic study on tones in Mandarin Chinese*. (Unpublished doctoral dissertation). Brown University, Providence, RI.

Tseng, C. Y., Pin, S. H., Lee, Y., Wang, H. M., & Chen, Y. C. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication*, *46*, 284–309.

Wang, H. S. (1998). An experimental study on the phonotactic constraints of Mandarin Chinese. In B. K. Tsou (Ed.), *Studia linguistica serica* (pp. 259–268). Hong Kong: Language Information Sciences Research Center at City University of Hong Kong.

Wang, X. (2013). Perception of Mandarin tones: The effect of L1 background and training. *Modern Language Journal*, *97*, 144–160.

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

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*, 3649–3658.

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning*, *54*, 681–712.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*, 25–47.

Wiener, S. (2015). *The representation, organization and access of lexical tone by native and non-native Mandarin speakers*. (Unpublished doctoral dissertation). The Ohio State University, Columbus, OH.

Wiener, S. (2017). Changes in early L2 cue-weighting of non-native speech: Evidence from learners of Mandarin Chinese. In *INTERSPEECH* (pp. 1765–1769). Stockholm, Sweden: International Speech Communication Association (ISCA).

Wiener, S. (2020). Second language learners develop non-native lexical processing biases. *Bilingualism: Language and Cognition*, *23*, 119–130.

Wiener, S., & Ito, K. (2015). Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition & Neuroscience*, *30*, 1048–1060.

Wiener, S., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, *56*, 38–51.

Wiener, S., Ito, K., & Speer, S. R. (2016). Individual variability in the distributional learning of L2 lexical tone. In J. Barnes, A. Brugos, S. Shattuck–Hufnagel, & N. Veilleux (Eds.), *Proceedings of the 8th International Conference on Speech Prosody* (pp. 538–542). Baixas, France: International Speech Communication Association.

Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 spoken word recognition combines input-based and knowledge-based processing. *Language and Speech*, *61*, 632–656.

Wiener, S., Lee, C. Y., & Tao, L. (2019). Statistical regularities affect the perception of second language speech: Evidence from adult classroom learners of Mandarin Chinese. *Language Learning*, *69*, 527–558.

Wiener, S., Murphy, T. K., Goel, A., Christel, M. G., & Holt, L. L. (2019). Incidental learning of non-speech auditory analogs scaffolds second language learners' perception and production of Mandarin lexical tones. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic sciences, Melbourne, Australia 2019* (pp. 1699–1703). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, *59*, 59–82.

Wong, P. (2013). Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan. *The Journal of the Acoustical Society of America*, *133*, 434–443.

Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native

English-speaking adults. *Applied Psycholinguistics*, *28*, 565–585.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*, 680–704.

Xing, J. Z. (2006). *Teaching and learning Chinese as a foreign language: A pedagogical grammar* (Vol. 1). Hong Kong: Hong Kong University Press.

Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. Berlin/Heidelberg: Springer.

Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice* (Vol. 1). Amsterdam: John Benjamins.

Zampini, M. L. (2008). L2 speech production research: Findings, issues, and advances. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 219–250). Amsterdam: John Benjamins.

Zhang, H. (2018). *Second language acquisition of Chinese tones: Beyond first-language transfer*. Leiden/Boston: Brill.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.