LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Incidental Nonspeech Auditory Learning Scaffolds Phonetic, Category, and Word Learning in a Foreign Language Classroom

Seth Wiener [a], Timothy K. Murphy,[b] and Lori L. Holt[c]

[a]Carnegie Mellon University [b]University of Wisconsin-Madison [c]University of Texas at Austin

Abstract: There is considerable lab-based evidence for successful incidental learning, in which a learner's attention is directed away from the to-be-learned stimulus and towards another stimulus. In this study, we extend incidental learning research into the language learning classroom. Three groups of adult second language (L2) learners ($N = 52$) engaged in structured classroom Mandarin learning took part in an 8-week study. One group served as a classroom-only control group. The second group underwent additional intentional auditory training involving Mandarin speech and explicit feedback. The third group underwent additional incidental learning combined with nonspeech "perceptual building block" categories—categories that share critical perceptual dimensions with target L2 speech categories but that are not perceived as speech.

CRediT author statement–**Seth Wiener**: conceptualization; methodology; investigation; formal analysis; visualization; validation; writing–original draft preparation; writing–review & editing; supervision; funding acquisition. **Timothy K. Murphy**: methodology; data curation; investigation; project administration; software; writing–review & editing. **Lori L. Holt**: conceptualization; methodology; investigation; writing–review & editing; supervision; funding acquisition.

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org

This work was supported by funding from The National Institutes of Health 1R03HD099382-01, The National Science Foundation BCS 2420979, and a *Language Learning* Early Career Research Grant. Christi L. Gomez, Atul Goel, and Youna Song provided critical support with methods and data collection. We thank Charlie Nagle and the anonymous reviewers for providing helpful, constructive feedback.

Correspondence concerning this article should be addressed to Seth Wiener, Department of Languages, Cultures & Applied Linguistics, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. Email: sethw1@cmu.edu

The handling editor for this article was Charlie Nagle.

DOI: 10.1111/lang.12700

We demonstrate that when supplemented with structured classroom learning, incidental learning involving nonspeech analogs promotes phonetic, category, and word learning equivalent to learning from more traditional intentional auditory training.

Keywords incidental learning; speech perception; second language acquisition; auditory category learning; lexical tone

Introduction

An adult second language (L2) learner faces at least two problems related to L2 speech perception. First, the learner must adapt a perceptual system attuned to first language (L1) input (Cutler, 2012). Second, a learner must receive enough L2 input to make meaningful new sound-to-meaning associations (Colantoni et al., 2015). To overcome these problems, researchers and instructors have traditionally drawn the learner's attention towards novel L2 speech sounds. Indeed, intentional auditory training (often combined with explicit feedback) can help learners build more robust and generalizable L2 speech categories (e.g., Bradlow et al., 1997; Chandrasekaran et al., 2016; Logan et al., 1991; Pisoni et al., 1982), though L1-like perceptual competence can take years, if not decades to achieve (e.g., Flege et al., 1999; Pelzl et al., 2021a; Richards, 2008).

We report on a classroom training study involving adult L2 incidental category learning in which a learner's attention is directed away from to-be-learned categories rather than towards them. We train learners on nonspeech "perceptual building block" categories—categories that share critical perceptual dimensions with target L2 speech categories but that are not perceived as speech. We demonstrate that when combined with more structured classroom learning, this type of incidental learning promotes phonetic, category, and word learning. These results provide exciting new directions for research on L2 sound learning.

Background Literature

Mandarin Tones as the Target Second Language Speech Sounds

Mandarin Chinese tones are characterized by their fundamental frequency (F0) over time (Chao, 1968; Ho, 1976) with secondary duration and amplitude cues (e.g., Zhang et al., 2022). Tone learning first involves a phonetic cue weighting problem. Mandarin-L1 English-L2 learners initially give prominent weight to F0 height, given its relative importance as an indexical cue in English (Chandrasekaran et al., 2010; Perrachione et al., 2011). Yet, F0 contour is more informative in Mandarin; weighting F0 height less and F0 contour more is the initial challenge for L2 learners (Liu et al., 2022; Wong & Perrachione, 2007). This shift in perception is difficult for most L2 learners, and the

literature commonly reports an initial period of improvement, followed by a lengthy learning plateau (Lee & Wiener, 2020). Hereafter, "phonetic learning" refers to F0 cue weighting.

Next, tone learning involves a category problem. Beijing Mandarin has four phonological tone categories: high-level (T1), rising (T2), low-dipping (T3), and high-falling (T4). Because F0 contours vary depending on the speaker and context (Xu, 1997), categories with similar starting F0 ranges (T2 and T3 both start with lower F0s; T1 and T4 both start with higher F0s) tend to be the most challenging for beginner L2 learners, though patterns of L2 acquisition can vary as a function of the task, stimuli, and participants (Hao, 2012, 2018; Tsukada et al., 2015; Yang, 2015). Typically, learners with musical training and better pitch perception abilities outperform those without musical training or weaker pitch perception abilities (e.g., Bent et al., 2006; Bowles et al., 2016; Lee & Hung, 2008; Li & DeKeyser, 2017), but large learner variability is common (Bramlett et al., 2024). Hereafter, "category learning" refers to categorizing tones into four categories defined by pitch movement.

Finally, tone learning involves a higher-level word problem. Tones, along with consonant and vowel segments, form morphemes and words: for example, T1 *jiē* "to pick up" and T4 *jiè* "to borrow." Recognizing one or more segments with a tone and matching that tone to some mental representation is typically the hardest step for L2 learners (e.g., Hao, 2024; Laméris & Post, 2022; Ling & Grüter, 2022). Pelzl et al. (2021a) have reported that even among Mandarin L2 learners with over a decade of experience (including L2 immersion), hundreds, if not thousands, of words are misremembered with incorrect tones. Hereafter, we use "word learning" to refer to this learning of higher-level information involving syllable and tone combinations mapped to a visual referent.

Intentional Second Language Tone Learning Within Structured Classrooms and Auditory Training

L2 Mandarin classroom instructors have traditionally been advised to explicitly draw the learner's attention towards tone contours (see Chan et al., 2022, and Xing, 2006, for reviews). This includes using the written romanization, such as *pīnyīn*, which marks tone contours on the vowels, instructors' hand gestures, colors, and other prominent visualizations of the contours (e.g., Farran & Morett, 2024; Godfroid et al., 2017; Morett, 2023; Morett & Chang, 2015; Ning et al., 2014; Yang, 2015). The four Mandarin tones are also explicitly taught with numeric labels and often as minimal pairs (e.g., T1 *jiē* and T4 *jiè*), which highlights the lexical function of tones. This approach ensures that learners are aware of the phonological categories and understand their general pitch patterns. For some adults, a 10-to-15-week structured Mandarin

language course results in small but measurable tone learning at the phonetic, category, and word levels (e.g., Hao, 2012; Liu & Wiener, 2020, 2022; Shen & Froud, 2016; Sun, 1997; Tsukada & Han, 2019; Wiener et al., 2019). For the majority of adults, however, it remains difficult to perceive these tone differences, particularly at the word level (e.g., Han & Tsukada, 2020; Wiener & Liu, 2021; Zou et al., 2017, 2022). The literature has reported slow learning gains, followed by a learning plateau, in which little to no progress has been made, with many learners abandoning the language learning process before reaching a high level of proficiency (e.g., Pelzl et al., 2021a, 2021b).

To overcome these challenges, L2 tone perception training research has traditionally focused on explicit, intentional auditory category learning with some form of instructional or corrective feedback (see Pelzl, 2019, for a review). Here, we use the term *intentional* to denote a deliberate, conscious effort, as the learner is aware of the target linguistic skills and knowledge being practiced. This intentional tone learning is often done individually (as compared to a group L2 classroom) and in a manner that provides individual or customized feedback, which may not always happen in a structured classroom. Intentional auditory training with feedback also allows for precise control of input and comparisons across stimuli, such as T1 *mā* and T2 *má* stimuli that differ only in F0 contour. Foundational research established behavioral differences on the basis of the listener's L1 and general linguistic experience with tonal and nontonal languages (Gottfried & Suiter, 1997; Hallé et al., 2004; Leather, 1987; Lee et al., 1996). This work showed that intentional lab-based auditory tone training can be effective for some participants—and even have long lasting effects (Wang et al., 1999)—but the literature has also reported that L2 learning varies from one individual to another depending on their L1 background, the stimuli used, and the training paradigm (Chen et al., 2023; Dong et al., 2019; Perrachione et al., 2011; Sadakata & McQueen, 2014; Wang, 2013; Yang, 2015).

Incidental Learning Combined With Nonspeech Analogs

We present an alternative to traditional intentional learning of tone: Incidental learning, in which a learner's attention is not directed towards the stimuli to be learned but rather towards what we believe is a potentially more engaging task distinct from the L2 learning domain. Here, we use the term *incidental* to denote learning of one stimulus while attending to another stimulus (see Schmidt, 1994), particularly when the learner is unaware of the target linguistic skills and knowledge being practiced. We build on an emerging body of cognitive science research that has demonstrated that incidental learning can lead to efficient auditory and speech category learning (Chan & Leung, 2020; Lim et al., 2013; Lim & Holt, 2011; Liu & Holt, 2011; Reber, 1989; Saito et al.,

2022; Seitz et al., 2010; Vlahou et al., 2012; Wade & Holt, 2005). Importantly, research in this domain has even shown that when incidental category learning takes place across nonlinguistic sounds, as it does in the current study, it engages neurobiological category-learning systems distinct from those involved in more explicit learning (e.g., Chandrasekaran et al., 2014; Gabay et al., 2015; Gabay & Holt, 2015; Lim et al., 2014, 2015, 2019). This incidental learning depends upon category exemplars that possess an underlying statistical regularity (Lim et al., 2019; Gabay et al., 2015; Wade & Holt, 2005) and the incidental alignment of this regularity with distinct behavioral outcomes in the primary task (Roark et al., 2022). This drives the posterior striatum to interact with left superior temporal sulcus (Lim et al., 2019)—a process implicated in complex auditory category perception and learning (Leech et al., 2009). In contrast, explicit training of difficult-to-acquire nonnative speech categories with overt, experimenter-provided feedback drives distinct patterns of striatal activation, centered more anteriorly and strongly linked to the presence of overt feedback (Tricomi et al., 2006). We extend this body of work to structured L2 learning, thus linking previous lab findings to the L2 classroom.

Our study makes use of a web-based space-themed videogame Neuraltone. The participants' explicit task is to navigate through the virtual world and target alien ships with a laser as the ships appear on screen (Figure 1, left). Players must shoot the alien spaceships before the ships reach the players' own spaceships, or else the players will lose life points. When a player is out of life points, the game will end. Players are not explicitly instructed to form audio-visual/audio-motor associations; they are not told about the significance of the sounds, and they do not overtly make sound categorizations. Two game features, however, strongly promote auditory category learning in an incidental manner. First, each spaceship is associated with a particular sound category. Each time a ship appears, a randomly-selected acoustically-variable sound exemplar drawn from an associated sound category is presented until the participant aims the laser and executes an action. These spaceships appear sequentially after the sound has begun playing. These nonspeech hums mimic the F0 contours typical of Mandarin tones (Liu, 2014; Obasih et al., 2023). As the game increases in difficulty, that is, reaches higher levels of gameplay, the speed at which alien spaceships appear also increases. Players must quickly adjust their laser as soon as the sounds begin playing in order to shoot the alien ships in time.

Second, each ship originates from a consistent quadrant of visual space (with jitter). Four categories are therefore established, each corresponding to the four sides of a monitor (Figure 1, right).

Figure 1 Neuraltone game play (left) and screen design (right).

Although not overtly required in Neuraltone, rapid sound categorization incidentally supports action. Upon hearing an alien sound, players must quickly position the laser towards one of the quadrants and begin firing in order to shoot the alien before it reaches the player's own spaceship. This encourages sound category learning through the categories' utility for functioning in the environment without requiring overt categorization responses or explicit feedback linked to such responses. This is not to suggest there is no feedback; listeners' predictions from the sounds lead to success or failure in the primary task. This, therefore, provides a form of incidental feedback, driven internally by the alignment of category exemplars possessing underlying statistical regularity and distinct behaviors in the primary task (Roark et al., 2022), which has been shown to engage distinct learning systems from explicit training tasks (Lim et al., 2019; Tricomi et al., 2006).

The Present Study: Incidental Learning of Second Language Tone Through Gameplay

Here we test whether incidental learning combined with nonspeech analogs can scaffold phonetic, category, and word learning in the L2 classroom. Specifically, we compare three groups of adults engaged in structured L2 Mandarin classes. The first group (+Classroom) serves as a classroom control group. This Classroom-only group did not take part in any additional training beyond their weekly class sessions (roughly 180 minutes per week) and outside preparation for class. The second group (+Classroom +Intentional Speech) took part in both classroom learning and weekly auditory training (roughly 30 min per week) using the Mandarin Acquisition Online (MAO) training program. The third group (+Classroom +Incidental Nonspeech) took part in both classroom learning and weekly Neuraltone game play (roughly 30 min per week). We used three behavioral tasks to tease apart contributions to phonetic, category, and word learning after roughly 8 weeks of training. Our research question was thus whether incidental learning combined with nonspeech analogs leads to greater tone learning outcomes as compared to traditional intentional auditory training with explicit feedback. Figure 2 summarizes the experimental design.

Method

All materials, data, and R code are available on the Open Science Framework (OSF). This study was carried out over roughly four years. Testing initially took place in person at the authors' university using the E-prime 2.0 software (Schneider et al., 2012). Roughly half (56%) of all data were collected in

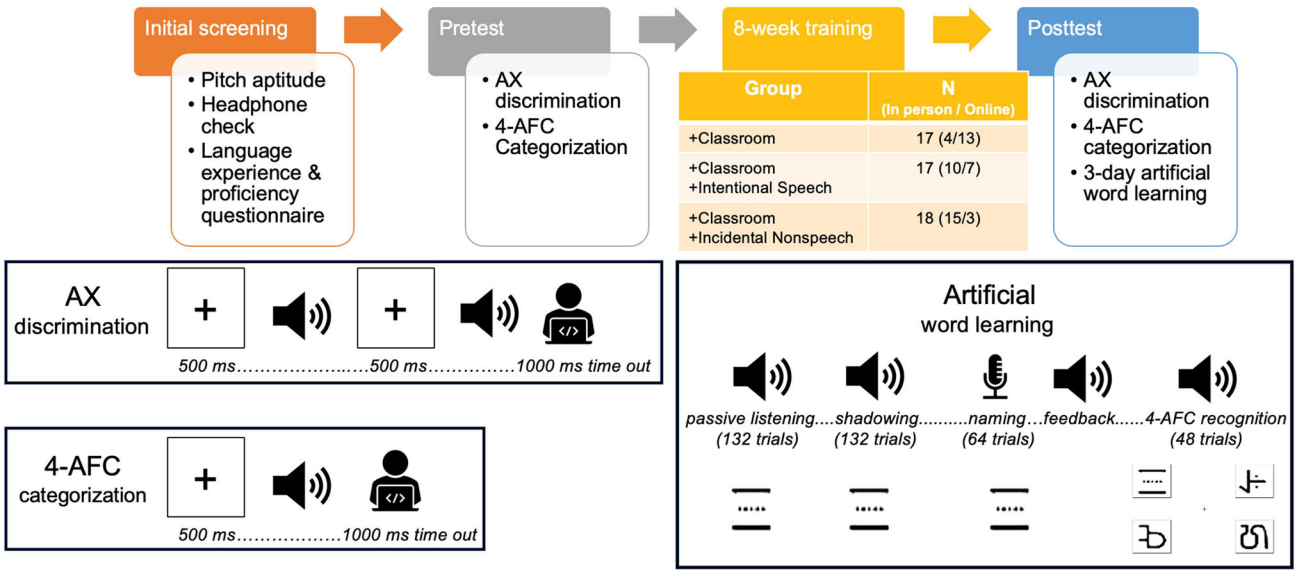

Figure 2 Study overview. Full experimental protocol shown at the top. Timing of the AX discrimination task (left middle) and the 4-AFC categorization task (left bottom) trials. Artificial word learning task protocol (right bottom). AX = same/different discrimination; 4-AFC = four-alternative-forced-choice.

person prior to the outbreak of the COVID-19 pandemic. In response to COVID-19, the pre- and posttest tasks were moved to the online platform Gorilla (Anwyl-Irvine et al., 2020). The remaining 44% of the data presented here were collected online. In other words, testing differed in software presentation and setting (E-prime in person / Gorilla online), but the stimuli and tasks were the same. Additional +Intentional Speech and +Incidental Nonspeech training outside the classroom were originally designed to be done via the internet on the participants' own time and were therefore identical for all participants. Whereas all participants engaged in L2 classroom learning used the same 15-week curriculum, syllabus, and textbooks and had the same four classroom instructors, participants who took part in the online testing also took part in remote, synchronous instruction via Zoom. This difference in learning modalities is an obvious limitation. See the Limitations Section for further discussion. The experiment was approved by the authors' Institutional Review Board.

Participants

All participants were 18 years of age or older with self-reported normal hearing and normal or corrected vison. In this paper, we report on 52 participants who self-reported their L1 as English ($N = 45$), German ($N = 1$), Korean ($N = 5$), and Spanish ($N = 1$). All participants were enrolled in a Mandarin language class and recruited through the class (either in person or online during the pandemic). Participants were pseudo-randomly assigned to one of the three groups in our study to ensure that each group had at least one participant from each classroom that we recruited from. Importantly, all participants in the two training groups completed at least 30 minutes of assigned training per week. An additional 19 L1 Mandarin participants from mainland China, who were living in the United States as university students at the time of testing, took part in all the tasks once in order to calculate an L1 mean for comparison. The mean age of all 71 participants was 20.8 years old, $SD = 2.7$. All participants were paid for their time.

Screening and Pretest

Screening ensured that participants were enrolled in a beginner Mandarin language class and fulfilled the requirements for the study. The Language Experience and Proficiency Questionnaire (Marian et al., 2007) was used to collect participant language histories. No participant self-identified as a heritage learner or had any prior Mandarin learning experience. Although any cutoff for years of music or voice training is arbitrary, we required that no participant should have more than five years of prior music or voice training

and no participant could be currently engaged in active music or voice learning (see Bramlett et al., 2024, and Wiener & Bradley, 2023, for additional music and tone discussions). Headphones were provided by the experimenter (in person) or confirmed via a short dichotic pitch test only perceivable when wearing headphones (Gorilla; Milne et al., 2021). A pitch aptitude task presented participants with two pure tones. Participants had to decide via button press whether the second tone was higher or lower in pitch than the first tone. The difference between the two tones started at 192 Hz and halved upon correct discrimination or doubled upon incorrect discrimination. A threshold was calculated through the adaptive maximum likelihood procedure (Green, 1993). All participants were able to reliably discriminate between two pure tones at 20 Hz or lower. A threshold of 20 Hz served to screen individuals with potential congenital amusia or tone deafness (Zhu et al., 2022) while retaining a sufficiently large and variable population. Equivalence tests were run following Lakens et al.'s (2018) study using half a standard deviation for Cohen's $d = 0.5$, as the smallest effect size of interest for pitch discrimination threshold. All three groups were statistically equivalent in their pitch discrimination behavior: $p = .25$, $p = .29$, $p = .09$.

In the same/different AX discrimination task, two speech sounds with a 500 ms interstimulus interval were played. Participants had to decide via button press whether the sounds A and X were the same or different as quickly and accurately as possible with a 1-second timeout period. Participants were given a short practice session, followed by 96 trials across two blocks. In each trial, the two sounds were always spoken by the same speaker. Trials were counterbalanced for an equal number of same/different and male/female trials. Stimuli creation followed Chandrasekaran et al.'s (2010) and Xu's (1997) designs: Two L1 Mandarin speakers (one female, one male) produced the Mandarin vowel /y/ with all four citation-form F0 contours. Praat's (Boersma & Weenink, 2015) pitch-synchronous overlap-add method was used to superimpose each F0 contour on the vowel. Stimuli were saved as 44.1 kHz wav files (16 bits) and had a normalized duration of 400 ms (female) and 450 ms (male).

In the four-alternative-forced-choice (4-AFC) categorization task, a speech sound was played. Participants had to decide via button press which of four Mandarin tone contours they perceived. Participants were told to respond as quickly and accurately as possible with a 1-second timeout period. The four Mandarin tone contours were shown on the screen as arrows indicating the F0 contours along with the numbers 1, 2, 3, 4 indicating which button mapped to the tone. Participants were given a short practice session, followed by 96 trials across two blocks. Trials were counterbalanced for an equal number of tone

type and male/female utterances. The stimuli were the same as those used in the AX task.

Training

Classroom. Participants were enrolled in a university Mandarin language class that met weekly for 180 minutes across three days. The course, which employed the communicative approach, was primarily structured around speaking and listening activities, and used a textbook developed by Wu et al. (2011). Students were assigned outside homework (roughly 60 min per day) and one-on-one practice with an L1 speaking assistant once a week (for roughly 30 min). Participants differed in their learning modality, with some participants meeting in person prior to the COVID-19 pandemic and others meeting online via Zoom (see Figure 2 for breakdown by group; see also the Limitations Section for further discussion).

Intentional Speech. After a participant completed their pretest tasks, a research assistant emailed the participant instructions for the online auditory training website MAO. This site hosted a series of common auditory tasks, such as same/different AX discrimination, ABX discrimination in which a listener had to determine whether sound X was the same as sound B or sound A, two-alternative-forced-choice (2-AFC) categorization task, 4-AFC categorization, and labeling of the syllable and tone. MAO stimuli involved speech from four different L1 speakers (two female; two male). Participants created an account, which tracked their MAO progress. Participants were told to complete 20 modules over the course of 8 weeks (roughly three modules per week). Given the design of MAO, participants were aware that MAO was tied to Mandarin segment and tone learning. Moreover, participants were explicitly told that the goal of additional MAO practice was to improve their perception of Mandarin speech. In other words, we attempted to make the learning as intentional as possible for the participants.

Each MAO module took approximately 20 minutes to complete, with modules increasing in number of speakers, difficulty of tasks, and the target speech sounds. For example, module 1 involved AX discrimination of one speaker producing different vowels in a syllable (e.g., *he* vs. *hui*) and different tones (e.g., *hé* vs. *hē*). Two speakers were introduced by module 8 along with more difficult tasks, and by module 20, four speakers were involved, and the tasks consisted only of labeling the correct syllable and tone. Figure 3 shows a screenshot of the MAO interface for an AX trial. Instant feedback (center), progress in the task (left), and a total correct/incorrect running score (right) were all shown to the user. Time on task was automatically logged for each participant.

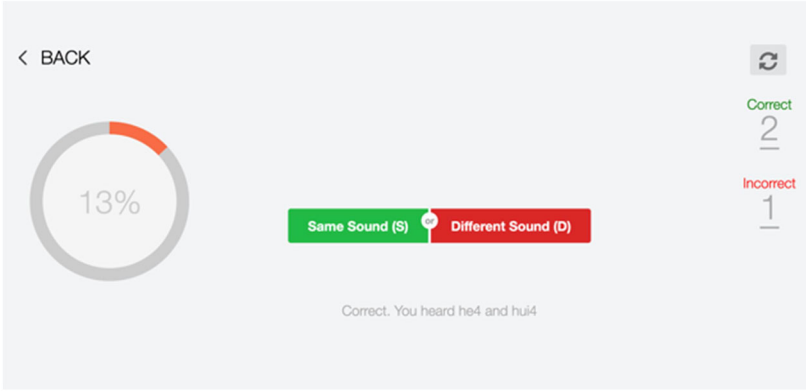

Figure 3 Mandarin Acquisition Online (MAO) auditory training interface for same/different AX discrimination trial.

Incidental Nonspeech. After a participant completed their pretest tasks, a research assistant emailed instructions to the participant for the online version of Neuraltone. Participants created an account, which logged gameplay metrics. Participants were told to play for roughly 1 hour a week in gameplay sessions lasting at least 20 minutes, for the next 8 weeks. There were no stipulations as to when the participants were to play the game during the week. Participants were neither told that the game was related to Mandarin nor that it could improve their Mandarin speech perception. In other words, unlike the +Intentional Speech group, the +Incidental Nonspeech group was naïve as to the goal of the training. After each gameplay session, participants were asked to report their score and total playtime for that session via Google forms. Time on task was recorded by the server hosting the game online, and the purpose of the self-report was primarily to keep participants aware of their progress throughout the week.

The +Incidental Nonspeech group's time on task playing the Neuraltone game (total minutes across the 8 weeks: $M = 256$, $SD = 8.8$ or an average of 32 minutes per week) was equivalent to the +Intentional Speech group's time on task using the MAO system (total minutes across the 8 weeks: $M = 260$, $SD = 6.9$ or an average of 32.5 minutes per week), $t(31.85) = -0.393$, $p = .65$ (Lakens et al., 2018; smallest effect size of interest: Cohen's $d = 0.5$).

Posttest

After a participant completed their training, a research assistant emailed the participant to set up the posttests across 3 consecutive days. On the first day,

participants completed the same AX discrimination and 4-AFC categorization tasks that they had done prior to training (counterbalanced in presentation order) and then began a 3-day artificial tonal language learning task. The stimuli and methods for the 3-day artificial tonal language learning task were identical to those used by Wiener et al. (2018, 2020, 2021)[1]. The stimuli consisted of 24 consonant-vowel syllables paired with one or more tone contours to create 82 words, with each word given a unique nonce symbol. These nonce words were phonotactically legal in standard Mandarin (analogous to the nonword blick in English). Each word was recorded by four different speakers (two female; two male) and saved at 44.4 kHz (16 bits). Each day involved a series of four tasks:

1. Self-paced passive learning task (132 trials). An image and its audio label were presented simultaneously, and the participant was asked to remember the pair.
2. Shadowing task (132 trials). An image–audio label pair was presented, and the participant was asked to repeat it aloud.
3. Naming task (64 trials). An image was shown, and the participant had to orally produce the audio label. Feedback was provided.
4. 4-AFC recognition task (48 trials). Four images were shown on the screen, and an audio label was played aloud. The participant was asked to click on the image that matched the audio label as quickly and accurately as possible. Feedback was provided.

Figure 2 (bottom right) shows an example using the item /ɹi/ with a falling tone across the first three tasks and then the 4-AFC trial (target: /ɹi/ with a falling tone, top left; unrelated distractor: /tsou̯/ with a dipping tone, top right; tonal competitor: /ɹi/ with a rising tone, bottom right; rhyme competitor: /fi/ with a falling tone, bottom left). Due to space constraints, we report only on the daily 4-AFC behavior.

Statistical Analyses

There are many ways to analyze our data[2]. In the interest of making our results as succinct and accessible as possible, we calculated an individual change score for each participant (i.e., pairwise difference) at the phonetic, category, and word levels. The AX discrimination data were used to calculate a by-participant d-prime (*d'*: hits–false alarms; Macmillan & Creelman, 2004) at the pretest and posttest. In other words, each participant's 96 data points at pretest were reduced to one *d'* sensitivity score; the same was done at posttest. The difference between the two scores was considered the change in phonetic

learning. The 4-AFC data were similarly used to calculate a by-participant total accuracy across all tone categories at pre- and posttest. The difference between the two tests was considered the change in category learning. The artificial language learning data were used to calculate a by-participant accuracy on day 1 and day 3 (thus ensuring sleep consolidation and lexical configuration; Liu & Wiener, 2020, 2022; Qin & Zhang, 2019). The difference between the two days was considered the change in word learning. From these data we built three linear regression models. Each model contained change in learning as the dependent variable and group as the dummy-coded independent variable, with the +Classroom group as the intercept or reference level. Coefficients could thus be interpreted in a straightforward manner representing increases and decreases in learning relative to the classroom-only group. In other words, by adding or subtracting the +Classroom +Incidental Nonspeech group's coefficient to/from the intercept, we arrived at that group's estimate or effect size. Models were releveled with +Classroom +Incidental Nonspeech as the reference level in order to obtain comparisons between the two training groups. Note that because each regression model had less than 60 observations, we do not report adjusted R^2 values, as they are considered unreliable (see Brysbaert, 2019, for a discussion). All data wrangling, visualization, and analyses were done using R (Version 4.3.1) along with multiple R packages (see OSF materials for further details on packages including version numbers).

Results

All Three Groups Showed Phonetic Learning

Figure 4 plots participants' *d'* results at pre- and posttests. Individual means (dots connected with gray line from pre- to posttest), group box and density plots, and the L1 mean (horizontal dotted line) are shown. This figure shows a general trend in which posttest *d'* was higher than pretest *d'*. For a small number of participants, posttest *d'* reached L1-like levels of sensitivity (+Classroom: 4 out of 17 participants; +Classroom +Intentional Speech: 3 out of 17 participants; +Classroom +Incidental Nonspeech: 5 out of 18 participants), but for the majority of participants, this was not the case. The regression model revealed a positive increase for the +Classroom group, $b = 1.33$, 95% CI [0.65, 2.01], $SE = 0.33$, $t = 3.93$, $p < .001$, and statistically similar positive increases for the +Classroom +Intentional Speech group, $b = 1.24$, 95% CI [0.55, 1.91], and +Classroom +Incidental Nonspeech group, $b = 1.33$, 95% CI [0.67, 1.99]. No difference was found between the two training groups, $p = .84$. All three groups, therefore, showed similar increases

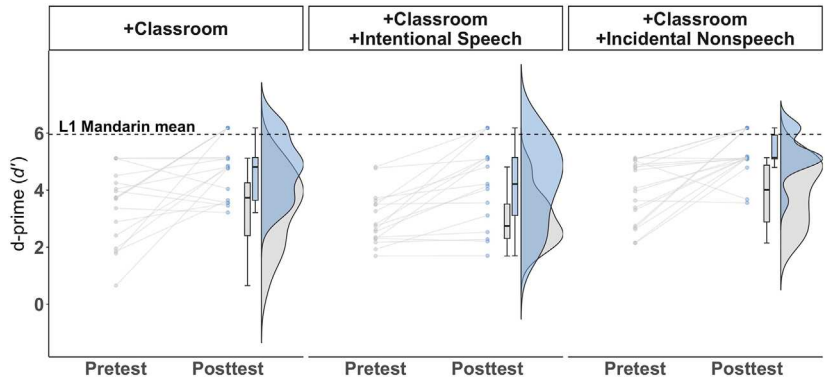

Figure 4 Change in sensitivity (same/different AX phonetic discrimination results). L1 = first language.

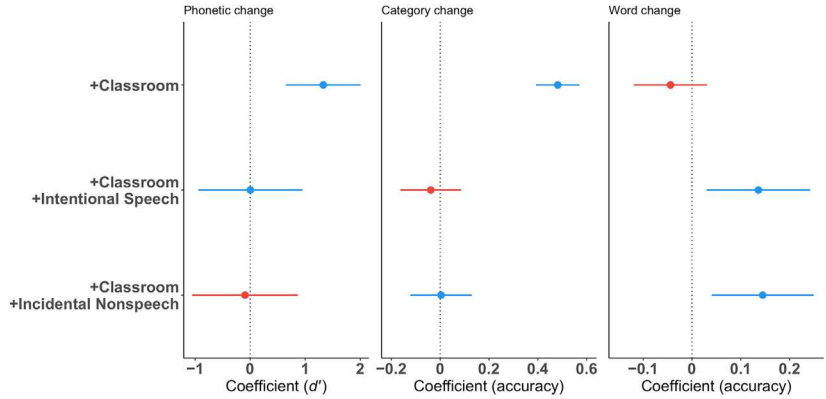

Figure 5 Linear regression model output showing phonetic (left), category (center), and word (right) coefficients. Dots represent model estimates, and whiskers show 95% confidence intervals. Blue indicates a positive coefficient. Red indicates a negative coefficient. The +Classroom group serves as the reference level or intercept. To arrive at each group's estimate, add the coefficient to the +Classroom intercept.

in sensitivity to tone as evidenced by the similar beta estimates and similar 95% confidence intervals.

Figure 5 (left panel) plots the model coefficients. The figure shows that the +Classroom group's estimate (and 95% confidence interval) does not cross 0, thus indicating a significant positive change. The two other groups' respective estimates cross 0, thus indicating that, relative to the +Classroom group, there was no additional increase.

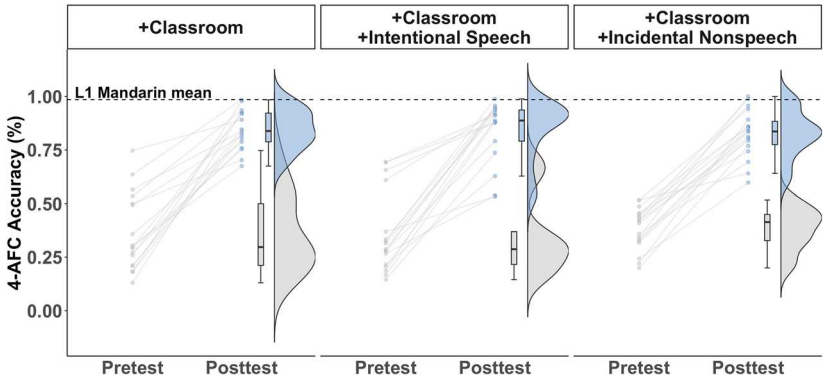

Figure 6 Change in tone categorization (4-AFC categorization). 4-AFC = four-alternative-forced-choice. L1 = first language.

All Three Groups Showed Category Learning

Figure 6 plots participants' 4-AFC results at pre- and posttests and uses the same plotting scheme as Figure 4. This plot shows relatively large increases for all groups. Once again, very few participants across all the groups approached L1-like behavior (+Classroom: 0 out of 17 participants; +Classroom +Intentional Speech: 1 out of 17 participants; +Classroom +Incidental Nonspeech: 2 out of 18 participants). The regression model (Figure 5, center panel) revealed a positive increase for the +Classroom group, $b = 0.48$, 95% CI [0.39, 0.57], $SE = 0.04$, $t = 10.87$, $p < .001$, and statistically similar positive increases for the +Classroom +Intentional Speech group, $b = 0.48$, 95% CI [0.40, 0.57], and +Classroom +Incidental Nonspeech group, $b = 0.44$, 95% CI [0.36, 0.53]. No difference was found between the two training groups, $p = .50$. Figure 5 shows that the two other groups' respective estimates cross 0, thus indicating that, relative to the +Classroom group, there was no additional increase. This can also be observed by noting the similar beta estimates and 95% confidence intervals for all three groups.

Only the +Intentional Speech and +Incidental Nonspeech Groups Showed Word Learning

Figure 7 plots participants' artificial tonal language learning 4-AFC results at days 1 and 3 using the same plotting scheme as Figures 4 and 6. This plot shows a wide range of learning outcomes at the word level, with many participants demonstrating little to no improvement from day 1 to day 3. No participants across the three groups approached L1-like word learning behavior[3]. The regression model (Figure 6, right panel) revealed no increase

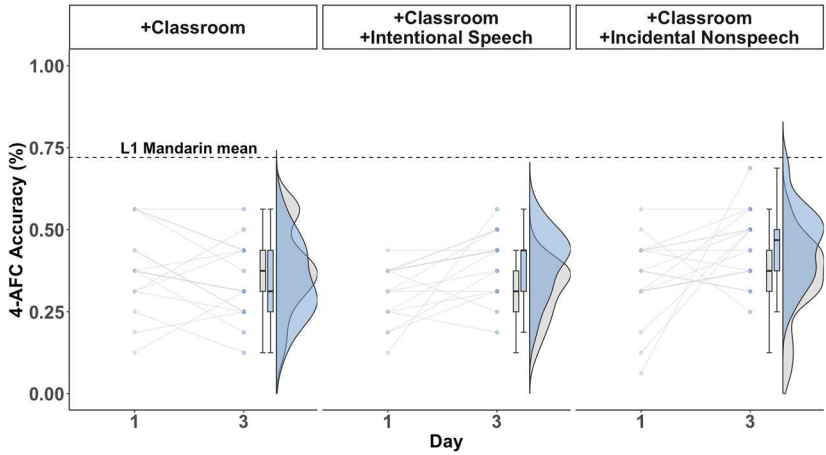

Figure 7 Change in artificial word learning (4-AFC categorization). 4-AFC = four-alternative-forced-choice. L1 = first language.

for the +Classroom group, $b = -0.04$, 95% CI [–0.12, 0.03], $SE = 0.04$, $t = -1.12$, $p = .24$, as the beta estimate crosses 0. In contrast, the two other groups show estimates that do not cross 0, that is, statistically significant positive increases for the +Classroom +Intentional Speech group, $b = 0.14$, 95% CI [0.02, 0.17], $SE = 0.05$, $t = 2.58$, $p = .01$, and the +Classroom +Incidental Nonspeech group, $b = 0.14$, 95% CI [0.03, 0.17], $SE = 0.05$, $t = 2.79$, $p = .007$. No difference was found between the two training groups, $p = .87$. Thus, only the two groups with additional outside training beyond the classroom showed improvement in word learning. This is particularly noteworthy for the +Incidental Nonspeech group, given that participants were trained on nonspeech analogs and were never told the purpose of the training, and yet still showed improved learning on novel speech.

Discussion

Adult L2 acquisition of Mandarin tone is a paradigmatic example of difficult to acquire speech categories. Much like L2 English acquisition of /r/ and /l/, L2 Mandarin acquisition of tone has been researched for decades using a wide range of methods and approaches (e.g., Leather, 1987; Hallé et al., 2004; Morett et al., 2022; Wang et al., 1999). In our 8-week study of L2 classroom learners, we used what we believed was a potentially more engaging videogame, in which tone-like nonspeech sounds served as gameplay cues. Players had to learn to associate distributions of nonspeech sounds in order

to advance to higher levels in the game. Incidental learning had been shown to improve categorization among naïve listeners in a lab setting (Lim & Holt, 2011; Wade & Holt, 2005), but our study was the first attempt to introduce incidental learning of nonspeech analogs to the L2 classroom. We specifically compared our incidental training approach to a more traditional intentional auditory training paradigm involving Mandarin speech samples with explicit feedback. We report three findings.

First, all three classroom groups demonstrated a significant increase in sensitivity to tone. The positive shift in d-prime was nearly identical for all three groups and demonstrated that structured L2 classroom learning improved phonetic cue weighting in line with previous studies (Hao, 2012; Wang et al., 2020). Interestingly, we observed no additive effects for the +Intentional Speech and +Incidental Nonspeech groups. One interpretation of these results is that for most beginners, there is an upper limit to how much phonetic learning can be reasonably captured across 8 weeks. It also remains an open question whether our population of learners will continue to improve or plateau, as is more commonly reported, after a first-year Mandarin course (Wiener, 2017).

Second, all three groups demonstrated a significant increase in tone categorization accuracy. The positive improvement was nearly identical across all three groups, which confirmed previous L2 tone category learning studies that had shown a large jump in accuracy given structured classroom input (Bramlett & Wiener, 2022; Shen & Froud, 2016; Wang et al., 2020; Wang, 2013). We note, however, that many of our learners still struggled to correctly categorize certain tones, especially tones 2 and 3, which are traditionally the hardest for beginner L2 learners given their overlapping F0 contours (e.g., Hao, 2012, 2018). Similar to our d-prime analysis, our 4-AFC analysis revealed no additive effects for the +Intentional Speech and +Incidental Nonspeech groups, suggesting that our classroom learners may have reached a ceiling. We tentatively conclude that for most of our L2 learners, both phonetic and category learning may have reached a limit on what was achievable over an 8-week training period.

Third, only the two classroom groups engaged in additional outside practice through +Intentional Speech training or +Incidental Nonspeech gameplay demonstrated a significant word learning improvement. Whereas the +Classroom group did not show a significant improvement in the word learning task from day 1 to day 3, the +Intentional Speech and +Incidental Nonspeech both showed nearly identical improvements, indicating that the additional practice did help participants better learn novel sound-to-image mappings involving tone. The failure of the +Classroom group to show a

significant improvement corroborates previous findings that forming robust phono-lexical connections remains the hardest challenge for most adult L2 learners (Hao, 2024; Laméris & Post, 2022; Ling & Grüter, 2022; Pelzl et al., 2021a, 2021b).

Whereas the connection of phonetic and category learning to the two training regimens is fairly straightforward, there is a lack of a similar direct connection between training and testing for word learning. These two training groups did not receive any additional word-specific training (at least not beyond what happened in the L2 classroom), and yet these groups showed improved novel word learning over multiple days. For the +Intentional Speech group, it could be that the additional Mandarin speech input from multiple speakers was beneficial for word learning, even if it was not fully absorbed as intake (see Rast, 2008; VanPatten, 2015). Indeed, limited evidence from adult L2 Mandarin noun (Liu & Wiener, 2020, 2022) and verb learning studies (Gao et al., 2022) supports a general usage-based account of L2 Mandarin lexical acquisition, in which even passive exposure to frequent sound patterns can lead to improved learning.

Our results, however, are even more surprising, as the +Incidental Nonspeech group did not hear additional Mandarin speech, and yet the nonspeech analogs still provided as much support as explicit training with feedback on real Mandarin speech. To the best of our knowledge, a traditional input/intake framework (e.g., Rast, 2008; VanPatten, 2015) or usage-based framework (e.g., Ellis & Wulff, 2014) does not consider nonspeech like the type we used in our study to be viable input/exemplars. One tentative conclusion that we can draw is that the nonspeech analogs that we used may be as effective at promoting word learning as other nonlinguistic cues like instructors' hand gestures, visual aids, and music (Farran & Morett, 2024; Morett, 2023; Morett & Chang, 2015; Ning et al., 2014; Wiener & Bradley, 2023). In each case, there is some overlapping dimension, which supports multimodal learning, dual coding, and embodied cognition (among other theories).

Thus, with respect to our research question concerning whether incidental learning combined with nonspeech analogs leads to greater tone learning outcomes than auditory training, for all three analyses, we found that the +Incidental Nonspeech group and +Intentional Speech group performed equally. This occurred despite the +Incidental Nonspeech group receiving input involving nonspeech analogs and playing a game that ostensibly had nothing to do with Mandarin language learning (and, as far as we know, none of the participants made this connection). In contrast, the +Intentional Speech group, well aware of the goal of the MAO training, were trained exclusively

on Mandarin speech and received constant feedback. Yet, we observed statistically similar phonetic, category, and word learning across the +Intentional Speech and +Incidental Nonspeech groups. When combined with structured L2 classroom learning, our incidental training routine together with nonspeech analogs was thus effective in promoting L2 speech learning at all three levels tested. Our results are in line with previous L2 incidental learning studies and highlight how such a training routine can lead to efficient auditory and speech category learning in the classroom (Chan & Leung, 2020; Lim et al., 2013, 2019; Lim & Holt, 2011; Reber, 1989; Saito et al., 2022; Seitz et al., 2010; Vlahou et al., 2012; Wade & Holt, 2005).

More importantly, our results open exciting new avenues for L2 speech sound research involving nonspeech analogs. We are in the process of examining whether these perception gains transfer to production and whether nonspeech analogs may help build robust L2 speech targets. Likewise, to what degree these results will transfer to more complicated stimuli, such as multitone sequences (i.e., disyllabic words), remains unknown. This avenue of research could prove especially fruitful, as the majority of spoken Mandarin words by type are disyllabic rather than monosyllabic, like the stimuli that we used.

Furthermore, varying the timing of incidental learning via gameplay could yield additional benefits. Gameplay with nonspeech analogs prior to structured L2 classroom learning may help scaffold even greater learning outcomes. Alternatively, delayed gameplay, in which L2 classroom learning occurs first, could also be highly effective, as this would allow learners to first acquire consonant–vowel syllable frames. Harder to acquire speech sounds—such as tones 2 and 3, which may take years to master (see Yang, 2015, for a discussion)—could be practiced through additional incidental training once the learner is more comfortable with Mandarin syllables.

Limitations and Conclusion

The COVID-19 pandemic undoubtably had an effect on our study. Potential confounds include the difference in classroom setting (in person vs. Zoom) and the increased psychological stress related to living through a global pandemic. We acknowledge that an online Zoom class is unable to provide students with the same type of interactions that an in-person classroom can. Available evidence suggests that during the pandemic, university students generally lacked motivation during online classes and perceived online classes to be less effective and less rigorous (e.g., Almahasees et al., 2021; Guillén et al., 2020). The online students in our study may have missed out on several in-person learning opportunities as well as dealt with additional pandemic

stress that made learning anything, especially Mandarin as a foreign language, more difficult, which in turn affected our results.

Moreover, the considerable variability that we observed in the word learning results is worth noting, though it is in line with previous studies (for discussions of individual differences in tone learning, see Perrachione et al., 2011; Sadakata & McQueen, 2014; Wong & Perrachione, 2007; Zou et al., 2017, 2022, among others). It is unclear whether some of this variation was due to COVID-19-related stress, but we believe that this variability highlights the need for future studies of individual differences, particularly the cognitive and linguistic variables that predict success in a language learning classroom (e.g., Bramlett et al., 2024; Chandrasekaran et al., 2010; Deng et al., 2016). We believe there is no "one-size-fits-all" approach to adult second language acquisition; some participants may benefit more from intentional learning strategies, such as our MAO training, whereas others may benefit more from incidental learning strategies like our Neuraltone game. It may also be that some of our participants responded positively to the gamification aspect of the task rather than its incidental learning aspect. It is known that gamified foreign language learning can be highly effective for some adults (Dubreil, 2020; Loewen et al., 2019). How to identify which learners best respond to which learning strategy remains an important open question for future research.

We also note that a handful of our participants (7 out of 52) were not L1 English speakers. Although none of these participants spoke another tonal language or had prior experience with Mandarin, it is theoretically possible that their additional experience with another L1 (Spanish, Korean, or German) may have played a role in their learning. To what degree this affected our results remains an open question.

It is also worth noting that we do not know to what degree the +Incidental Nonspeech group knew that the Neuraltone game was aimed at improving their Mandarin speech perception. It may be that several participants understood that the gameplay sounds mimicked Mandarin tones (it was called Neuraltone, after all), whereas some may have been oblivious to the purpose of the sounds during gameplay. Whether this awareness of the game's goals affected our results remains unclear and an important direction for future research in this area.

Regarding our analyses, we acknowledge that all three of our statistical models had relatively low statistical power given our sample size ($N = 52$). Our online R code includes several post hoc exploratory analyses on subsets of data. In these analyses, we see the same general trends as those that we report here, which suggests that our results would likely be the same with a more tightly controlled study (though this remains an empirical question). Across

all three of our analyses, we reported relatively large beta estimates, which strengthens our claim that incidental learning is effective when combined with classroom learning. Importantly, this result was observed even in spite of the large individual variability that is commonly seen in L2 speech research. We welcome future research that makes use of our open methods and materials to carry out a close or partial replication.

Finally, it is worth asking whether the additional time that our participants spent outside of class devoted to Neuraltone training might be better used in some other way. Would 30 minutes of focused word learning with course materials yield similar results? Would using an app like Duolingo yield similar results? This remains an empirical question. Whereas we did not carry out a formal interview or postgameplay survey, the informal feedback that we received from the participants who played the game was overwhelmingly positive. Participants reported that the game was "fun," "easy to play," and "a nice break from studying." We believe that the additional 30 minutes of weekly Neuraltone gameplay is worth the investment, given the results we present here.

To conclude, our data support the claim that incidental learning combined with nonspeech auditory analogs scaffold learning of complex, novel categories that share critical perceptual dimensions with target L2 speech categories. Moreover, when combined with structured L2 classroom learning, incidental learning is as effective as traditional auditory training involving speech while simultaneously providing the learner with what we believe is a potentially more engaging form of practice outside the classroom.

Final revised version accepted 16 November 2024

Notes

1 A 4th day of testing involving eye-tracking was initially planned but was removed due to COVID-19.
2 See online R code for additional approaches, including paired *t*-tests and mixed-effects models.
3 One participant in the +Classroom +Incidental Nonspeech group did, however, come close to L1-like word learning behavior. This participant also showed the most improvement, having scored the lowest accuracy on day 1 and the highest accuracy on day 3.

References

Almahasees, Z., Mohsen, K., & Amin, M. O. (2021). Faculty's and students' perceptions of online learning during COVID-19. *Frontiers in Education*, *6*, Article 638470. https://doi.org/10.3389/feduc.2021.638470

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-407. https://doi.org/10.3758/s13428-019-01237-x

Bent, T., Bradlow, A. R., & Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human perception and performance*, *32*(1), 97–103. https://doi.org/10.1037/0096-1523.32.1.97

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer (Version 5.3.62) [Computer software] http://www.praat.org/

Bowles, A.R., Chang, C.B., & Karuzis, V.P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*(4), 774–808. https://doi.org/10.1111/lang.12159

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310. https://doi.org/10.1121/1.418276

Bramlett, A. A., Brown, B., Dueck, J., & Wiener, S. (2024). Measuring music and prosody: accounting for variation in non-native speech discrimination with working memory, specialized music skills, and music background. *Speech Prosody*, 482–486. https://doi.org/10.21437/SpeechProsody.2024-98

Bramlett, A. A., & Wiener, S. (2022). jTRACE modeling of L2 Mandarin learners' spoken word recognition at two time points in learning. *Speech Prosody*, 773–776. https://doi.org/10.21437/SpeechProsody.2022-157

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Chan, J., Woore, R., Molway, L., & Mutton, T. (2022). Learning and teaching Chinese as a foreign language: A scoping review. *Review of Education*, *10*(3), Article e3370. https://doi.org/10.1002/rev3.3370

Chan, R. K., & Leung, J. H. (2020). Why are lexical tones difficult to learn?: Insights from the incidental learning of tone-segment connections. *Studies in Second Language Acquisition*, *42*(1), 33–59. https://doi.org/10.1017/S0272263119000482

Chandrasekaran, B., Sampath, P.D., & Wong, P. C. M. (2010). Individual variability in cueweighting and lexical tone learning. *Journal of the Acoustical Society of America*, *128*(1), 456–465. https://doi.org/10.1121/1.3445785

Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488–495. https://doi.org/10.3758/s13423-013-0501-5

Chandrasekaran, B., Yi, H. G., Smayda, K. E., & Maddox, W. T. (2016). Effect of explicit dimensional instruction on speech category learning. *Attention, Perception, & Psychophysics*, *78*(2), 566–582. https://doi.org/10.3758/s13414-015-0999-x

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chen, J., Antoniou, M., & Best, C. T. (2023). Phonological and phonetic contributions to perception of non-native lexical tones by tone language listeners: Effects of memory load and stimulus variability. *Journal of Phonetics*, *96*, Article 101199. https://doi.org/10.1016/j.wocn.2022.101199

Colantoni, L., Steele, J., & Escudero, P. (2015). *Second language speech*. Cambridge University Press.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press. https://doi.org/10.7551/mitpress/9012.001.0001

Deng, Z., Chandrasekaran, B., Wang, S., & Wong, P. C. (2016). Resting-state low-frequency fluctuations reflect individual differences in spoken language learning. *Cortex*, *76*, 63–78. https://doi.org/10.1016/j.cortex.2015.11.020

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. *PeerJ*, *7*, Article e7191. https://doi.org/10.7717/peerj.7191

Dubreil, S. (2020). Using games for language learning in the age of social distancing. *Foreign Language Annals*, *53*(2), 250–259. http://doi.org/10.1111/flan.12465

Ellis, N. C., & Wulff, S. (2014). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 87–105). Routledge.

Farran, B. M., & Morett, L. M. (2024). Multimodal cues in L2 lexical tone acquisition: Current research and future directions. *Frontiers in Education*, *9*, Article 1410795. https://doi.org/10.3389/feduc.2024.1410795

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*(1), 78–104. https://doi.org/10.1006/jmla.1999.2638

Gabay, Y, Dick, F. K., Zevin, J., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1124–1138. https://doi.org/10.1037/xhp0000073

Gabay, Y., & Holt, L. L. (2015). Incidental learning of sound categories is impaired in developmental dyslexia. *Cortex*, *73*, 131–143. https://doi.org/10.1016/j.cortex.2015.08.008

Gao, Z., Wiener, S., & MacWhinney, B. (2022). Acquisition of Chinese verb separation by adult L2 learners. *Languages*, *7*(3), Article 225. https://doi.org/10.3390/languages7030225

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*(4), 819–857. https://doi.org/10.1111/lang.12246

Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, *25*(2), 207–231. https://doi.org/10.1006/jpho.1997.0042

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes–no task. *The Journal of the Acoustical Society of America*, *93*(4), 2096–2105. https://doi.org/10.1121/1.406696

Guillén, G., Sawin, T., & Avineri, N. (2020). Zooming out of the crisis: Language and human collaboration. *Foreign Language Annals*, *53*(2), 320–328. https://doi.org/10.1111/flan.12459

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*(3), 395–421. https://doi.org/10.1016/S0095-4470(03)00016-0

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

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

Hao, Y. C. (2018). Second language perception of Mandarin vowels and tones. *Language and Speech*, *61*(1), 135–152. https://doi.org/10.1177/0023830917717759

Hao, Y. C. (2024). Phonolexical processing of Mandarin segments and tones by English speakers at different Mandarin proficiency levels. *Second Language Research*, *40*(3), 533–557. https://doi.org/10.1177/02676583231167790

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, *33*(5), 353–367. https://doi.org/10.1159/000259792

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Laméris, T. J., & Post, B. (2022). The combined effects of L1-specific and extralinguistic factors on individual performance in a tone categorization and word identification task by English-L1 and Mandarin-L1 speakers. *Second Language Research*, *39*(3). 833–871. https://doi.org/10.1177/02676583221090068

Leather, J. (1987). F0 pattern inference in the perceptual acquisition of second language tone. In A. James & J. Leather (Eds.), *Sound patterns in second language acquisition* (pp. 59–80). De Gruyter Mouton. https://doi.org/10.1515/9783110878486-005

Lee, C. Y., & Hung, T. H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *The Journal of the Acoustical Society of America*, *124*(5), 3235–3248. https://doi.org/10.1121/1.2990713

Lee, C. Y., & Wiener, S. (2020). Acoustic-based and knowledge-based processing of Mandarin tones by native and non-native speakers. In H. Liu, F. Tsao & P. Li (Eds.), *Speech perception, production and acquisition. Multidisciplinary approaches in Chinese languages* (pp. 37–57). Springer. 10.1007/978-981-15-7606-5_3

Lee, Y.-S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: a cross-linguistic comparison. *Journal of Psycholinguistic Research*, *25*(5), 527–542. https://doi.org/10.1007/BF01758181

Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *Journal of*

Neuroscience, *29*(16), 5234–5239. https://doi.org/10.1523/JNEUROSCI.5758-08.2009

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*(4), 593-620. https://doi.org/10.1017/S0272263116000358

Lim, S. J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in Neuroscience*, *8*, Article 230. https://doi.org/10.3389/fnins.2014.00230

Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences*, *116*(10), 4671–4680. https://doi.org/10.1073/pnas.1811992116

Lim, S-J., Fiez, J. A., Wheeler, M. E., & Holt, L. L. (2013, April 13–16). *Investigating the neural basis of video-game-based category learning*. [Paper presentation] 20th Anniversary Meeting of the Cognitive Neuroscience Society, San Francisco, CA, United States.

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lim, S. J., Lacerda, F., & Holt, L. L. (2015). Discovering functional units in continuous speech. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1139–1152. https://doi.org/10.1037/xhp0000067

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

Liu, J., & Wiener, S. (2020). Homophones facilitate lexical development in a second language. *System*, *91*, Article 102249. https://doi.org/10.1016/j.system.2020.102249

Liu, J., & Wiener, S. (2022). Effects of phonological and talker familiarity on second language lexical development. *The Mental Lexicon*, *17*(1), 132–153. https://doi.org/10.1075/ml.20024.liu

Liu, L., Yuan, C., Ong, J. H., Tuninetti, A., Antoniou, M., Cutler, A., & Escudero, P. (2022). Learning to perceive non-native tones via distributional training: Effects of task and acoustic cue weighting. *Brain Sciences*, *12*(5), Article 559. https://doi.org/10.3390/brainsci12050559

Liu, R. (2014). *Category learning supporting non-native speech perception: Investigating issues of variability, generalization, and transfer*. [Doctoral dissertation, Carnegie Mellon University]. https://doi.org/10.1016/j.cognition.2023.105467

Liu, R., & Holt, L. L. (2011). Neural changes associated with non-speech auditory category learning parallel those of speech category acquisition. *Journal of Cognitive Neuroscience*, *23*(3), 683–698. https://doi.org/10.1162/jocn.2009.21392

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311. https://doi.org/10.1017/S0958344019000065

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and /l/: a first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press. https://doi.org/10.4324/9781410611147

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language & Hearing Research*, *50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Morett, L. M. (2023). Observing gesture at learning enhances subsequent phonological and semantic processing of L2 words: An N400 study. *Brain and Language*, *246*, Article 105327. https://doi.org/10.1016/j.bandl.2023.105327

Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105

Morett, L. M., Feiler, J. B., & Getz, L. M. (2022). Elucidating the influences of embodiment and conceptual metaphor on lexical and non-speech tone learning *Cognition*, *222*, Article 105014. https://doi.org/10.1016/j.cognition.2022.105014

Ning, L. H., Shih, C., & Loucks, T. M. (2014). Mandarin tone learning in L2 adults: A test of perceptual and sensorimotor contributions. *Speech Communication*, *63*, 55–69. https://doi.org/10.1016/j.specom.2014.05.001

Obasih, C. O., Luthra, S., Dick, F., & Holt, L. L. (2023). Auditory category learning is robust across training regimes. *Cognition*, *237*, Article 105467. https://doi.org/10.1016/j.cognition.2023.105467

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. The Journal of the Chinese Language Teachers Association, USA*, *54*(1), 51–78. https://doi.org/10.1075/csl.18009.pel

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 potentials 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*, Article 689423. https://doi.org/10.3389/fcomm.2021.689423

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (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*(1), 461–472. https://doi.org/10.1121/1.3593366

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 297–314. https://doi.org/10.1037/0096-1523.8.2.297

Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PLOS ONE*, *14*(12), Article e0221498. https://doi.org/10.1371/journal.pone.0221498

Rast, R. (2008). *Foreign language input: Initial processing* (Vol. *28*). Multilingual Matters.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Richards, J. C. (2008). *Moving beyond the plateau: From intermediate to advanced levels in language learning*. Cambridge University Press.

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The representational glue for incidental category learning is alignment with task-relevant behavior. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(6), 769–784. https://doi.org/10.1037/xlm0001078

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*, Article 1318. https://doi.org/10.3389/fpsyg.2014.01318

Saito, K., Hanzawa, K., Petrova, K., Kachlicka, M., Suzukida, Y., & Tierney, A. (2022). Incidental and multimodal high variability phonetic training: Potential, limits, and future directions. *Language Learning*, *72*(4), 1049–1091. https://doi.org/10.1111/lang.12503

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165–209). Academic Press.

Schneider, W., Eschman, A., and Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools, Inc.

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, *115*(3), 435–443. https://doi.org/10.1016/j.cognition.2010.03.004

Shen, G., & Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America*, *140*(6), 4396–4403. https://doi.org/10.1121/1.4971765

Sun, S. H. (1997). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese*. Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*(6), 1029–1043. https://doi.org/10.1162/jocn.2006.18.6.1029

Tsukada, K. & Han J. I. (2019). The perception of Mandarin lexical tones by native Korean speakers differing in their experience with Mandarin. *Second Language Research*, *35*(3), 305–318. https://doi.org/10.1177/0267658318775155

Tsukada, K., Xu, H. L., & Rattanasone, N. X. (2015). The perception of Mandarin lexical tones by listeners from different linguistic backgrounds. *Chinese as a Second Language Research*, *4*(2), 141–161. https://doi.org/10.1515/caslar-2015-0009

VanPatten, B. (2015). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 113–134). Routledge.

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*(2), 363–381. https://doi.org/10.1037/a0025014

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, *118*(4), 2618–2633. https://doi.org/10.1121/1.2011156

Wang, T., Potter, C. E., & Saffran, J. R. (2020). Plasticity in second language learning: The case of Mandarin tones. *Language Learning and Development*, *16*(3), 231–243. https://doi.org/10.1080/15475441.2020.1737072

Wang, X. (2013). Perception of Mandarin tones: The effect of L1 background and training. *The Modern Language Journal*, *97*(1), 144–160 https://doi.org/10.2307/23361743

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*(6), 3649–3658. https://doi.org/10.1121/1.428217

Wiener, S. (2017). Changes in early L2 cue-weighting of non-native speech: Evidence from learners of Mandarin Chinese. *INTERSPEECH*, 1765–1769. https://doi.org/10.21437/Interspeech.2017-289

Wiener, S., & Bradley, E. D. (2023). Harnessing the musician advantage: Short-term musical training affects non-native cue weighting of linguistic pitch. *Language Teaching Research*, *27*(4), 1016–1031. https://doi.org/10.1177/1362168820971791

Wiener, S., Chan, M. K., & Ito, K. (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions? *The Modern Language Journal*, *104*(1), 152–168. https://doi.org/10.1111/modl.12619

Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 spoken word recognition combines input-based and knowledge-based processing. *Language and Speech*, *61*(4), 632–656. https://doi.org/10.1177/0023830918761762

Wiener, S., Ito, K., & Speer, S. R. (2021). Effects of multitalker input and instructional method on the dimension-based statistical learning of syllable-tone combinations: An eye-tracking study. *Studies in Second Language Acquisition*, *43*(1), 155–180. https://doi.org/10.1017/S0272263120000418

Wiener, S., & Lee, C. Y. (2020). Multi-talker speech promotes greater knowledge-based spoken Mandarin word recognition in first and second language listeners. *Frontiers in Psychology*, *11*, Article 214. https://doi.org/10.3389/fpsyg.2020.00214

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*(3), 527–558. https://doi.org/10.1111/lang.12342

Wiener, S., & Liu, J. (2021). Effects of perceptual abilities and lexical knowledge on the phonetic categorization of second language speech. *JASA Express Letters*, *1*(4), Article 045202. https://doi.org/10.1121/10.0004259

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. https://doi.org/10.1017/S0142716407070312

Wu, S. M., Yu, Y., Zhang, Y., & Tian, W. (2011). *Chinese link (level 1)*. Pearson.

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

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61–83. https://doi.org/10.1006/jpho.1996.0034

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

Zhang, H., Wiener, S., & Holt, L. L. (2022). Adjustment of cue weighting in speech by speakers and listeners: Evidence from amplitude and duration modifications of Mandarin Chinese tone. *The Journal of the Acoustical Society of America*, *151*(2), 992–1005. https://doi.org/10.1121/10.0009378

Zhu, J., Chen, X., Chen, F., & Wiener, S. (2022). Individuals with congenital amusia show degraded speech perception but preserved statistical learning for tone languages. *Journal of Speech, Language, and Hearing Research*, *65*(1), 53–69. https://doi.org/10.1044/2021_JSLHR-21-00383

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

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*, Article 891756. https://doi.org/10.3389/fpsyg.2022.891756

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary