To be published in Studies in *Second Language Acquisition* (Cambridge University Press)

A Longitudinal Investigation of Explicit and Implicit Auditory Processing in L2 Segmental and Suprasegmental Acquisition

Hui Sun[1]

Kazuya Saito

Adam Tierney

Abstract

Precise auditory perception at a subcortical level (neural representation and encoding of sound) has been suggested as a form of implicit L2 aptitude in naturalistic settings. Emerging evidence suggests that such implicit aptitude explains some variance in L2 speech perception and production among adult learners with different first language backgrounds and immersion experience. By examining 46 Chinese learners of English, the current study longitudinally investigated the extent to which explicit and implicit auditory processing ability could predict L2 segmental and prosody acquisition over a five-month early immersion. According to the results, participants' L2 gains were associated with more explicit and integrative auditory processing ability (remembering and reproducing music sequences), while the role of implicit, preconscious perception appeared to be negligible at the initial stage of post-pubertal L2 speech learning.

[1] The project was funded by the Birkbeck Global Challenges Research Fund "*Impaired auditory processing as a cause of difficulty with second language learning in native Mandarin speakers*" (awarded to AT), and the Leverhulme Trust Research Grant "*Does having a good ear promote successful second language speech learning?*" (awarded to KS and AT).

Introduction

It is widely acknowledged that the great individual variability in post-puberty second language (L2) learning success cannot be solely explained by experience factors (i.e., the extent to which learners practice a target language). This is arguably because even when exposed to a similar amount and quality of L2 input, learners differ in the receptive and productive L2 skills they can achieve. One important factor contributing to these individual differences could be perceptual (e.g., phonemic discrimination) and broader cognitive abilities (e.g., working memory) that are involved in the underlying mechanisms of language processing. Variability in these perceptual and broader cognitive foundations, therefore, could help determine an individual's readiness to learn a language (i.e., language learning aptitude)—and therefore, measuring these skills could enable a further examination of the association between aptitude and second language learning gains which is the goal of the current study.

Second Language Learning Aptitude

Over the past five decades, the role of aptitude in second language acquisition (SLA) has been extensively researched (see Li, 2016 for a review). Originally, aptitude was conceptualized as perceptual and broader cognitive abilities that help determine the success of intentional and explicit L2 learning in classroom settings (i.e., foreign language aptitude). According to the seminal model proposed by Carroll and Sapon (1959), the components of aptitude include phonetic coding, grammatical sensitivity, inductive learning, and rote memory. These abilities are believed to be essential to the explicit processing stages of noticing, patterning and associating (Li, 2015; Skehan, 2002). Indeed, measured by the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1959), aptitude scores were found to demonstrate moderate correlations with foreign language achievement, especially at the initial learning stage (e.g., Bialystok & Fröhlich, 1978; Sparks et al., 1998). Similar findings have been reported using other widely used aptitude tests based on Carroll's model such as the Pimsleur Language Aptitude Battery (PLAB; Pimsleur, 1966) and the LLAMA Test (Meara, 2005). For example, by relating PLAB aptitude scores to the final foreign language grades of high school students, Curtin, Avner

and Smith (1983) found that more variance was explained by aptitude among beginner-level students than among advanced-level students.

More recently, researchers have begun to examine the relationship between aptitude and L2 learning in naturalistic settings. In the existing literature, the explanatory power of traditional foreign language aptitude varies according to a range of factors, such as age of onset and the type of linguistic structures being learned or assessed (e.g., Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; Granena & Long, 2013). In light of the more complex nature of naturalistic L2 learning and processing, a growing number of scholars have emphasized the importance of capturing one's ability to learn a language not only explicitly (i.e., with conscious awareness) but also implicitly (i.e., without conscious awareness) (Doughty, 2019). Thus, in the current study, the former type of ability is labeled as explicit language aptitude and the latter type is labeled as implicit language aptitude.

To capture the language aptitude of both types, a range of instruments have recently been introduced to the SLA field. For example, the Hi-LAB test battery (Linck et al., 2013) was designed to identify the perceptual and broader cognitive abilities that could predict adult learners' L2 listening and reading attainment at a highly advanced level. Out of the eleven domain-general cognitive abilities (underlying general-purpose learning) and domain-specific perceptual abilities (specific to language learning) covered in Hi-LAB, implicit sequence learning ability (measured via serial reaction time), associative memory, and phonological short-term memory were found to be predictors of high-level L2 achievement, whereas the two measures of domain-specific perceptual abilities (phonemic discrimination and categorization) were not.

One domain-general perceptual ability, however, has been largely overlooked by previous investigations of language learning aptitude but nonetheless could play an important role in language acquisition—auditory processing. This refers to the ability to precisely and accurately perceive sound characteristics, which is commonly assessed behaviorally by asking participants to discriminate or reproduce individual acoustic dimensions of sound, or by examining the size, stability, or accuracy of neural responses to sound via neurophysiological measures.

Auditory Processing Ability and L1 Acquisition in Children and Adults

Speech contains information about language structure on many different levels, from acoustic patterns on a time scale of tens of milliseconds which distinguish phonemes, to patterns unfolding over seconds which convey information about conversational turn-taking, with many levels in between. Variation along several acoustic dimensions, including pitch, duration, amplitude, and spectral shape (formants), conveys information about many different aspects of language, including word boundaries (Cutler & Butterfield, 1992), lexical stress (Fear, Cutler, & Butterfield, 1995), phrase boundaries (Marslen-Wilson, Tyler, Warren, Grenier, & Lee, 1992), and information structure (Breen, Fedorenko, Wagner, & Gibson, 2010). Robust auditory processing may be facilitated by decreased variability of perceptual input and lead to rapid acquisition of knowledge about phonemic and prosodic categories (Toscano & McMurray, 2010), with potential beneficial consequences for the learning of language structures on multiple levels.

There is a long history of research in the L1 acquisition literature on the relationship between precise auditory processing and various language skills, including reading and grammatical knowledge. In this literature, auditory processing has been measured in many different ways; here we will focus on three particularly popular techniques. First, the degree of noise in a particular auditory channel can be measured using adaptive discrimination tests (psychoacoustic thresholds). For example, children with specific language impairment (SLI) or dyslexia have been shown to be more likely to have higher thresholds along a number of different auditory dimensions, including frequency (McArthur & Bishop, 2005), duration (Casini, Pech-Georgel, & Ziegler, 2018), and amplitude rise time (Goswami et al., 2002). In the same vein, adults with psychoacoustic difficulties are more likely to have reading difficulties (Ahissar, Protopapas, Reid, & Merzenich, 2000; Walker, Shinn, Cranford, & Givens, 2002). Second, the ability to discriminate or reproduce melodic and rhythmic patterns can be assessed, and these abilities tend to be somewhat poorer in children with poor phonological and reading skills (Flaugnacco et al., 2014; Grube, Kumar, Cooper, Turton, & Griffiths, 2012).

The advantage of these behavioral techniques is that they are relatively cost-effective, quick to implement, and simple enough to be performed by children. However, a major disadvantage of these techniques is that they touch on only conscious and attentional processing

of sounds (rather than bottom-up, implicit auditory processing) and could also reflect modality-general (i.e. not specific to sound) cognitive skills such as attention (Snowling, Gooch, McArthur, & Hulme, 2018). One way to attempt to isolate implicit auditory processing is to make use of neural measures of sound processing which are comparatively unaffected by cognitive state. One such measure is an electroencephalographic (EEG) response known as the frequency following response (FFR), which is the third way to assess auditory processing.

The FFR was first recorded in 1973 by Moushegian, Rupert and Stillman. They reported that periodic auditory stimuli give rise to an EEG response that mirrors the frequency content of the stimulus—i.e. a pure tone repeating at 400 cycles per second would give rise to a roughly sinusoidal response repeating 400 times a second. The fidelity of the response to the stimulus is such that the difference between vowel sounds presented to participants is identifiable based on the difference in the frequency content of the response (Galbraith, Bhuta, Choate, Kitahara, & Mullen, 1998). The FFR is a rapid response, beginning only 10 milliseconds after a sound is presented to a participant. This suggests that its primary generator is not the cerebral cortex, given that 10ms is not sufficient for sound information to reach the cortex. Indeed, early work suggested that the FFR is produced exclusively by sub-cortical generators. For example, Smith, Marsh and Brown (1975) showed that cooling of the inferior colliculus (a region in the auditory midbrain) greatly diminished the scalp-recorded FFR, while Sohmer, Pratt and Kinarti (1997) showed that patterns with upper brain-stem lesions show no scalp-recorded FFR. Moreover, Kiren, Aoyagi, Furuse and Koike (1994) showed that lesioning the inferior colliculus in cats greatly diminishes the FFR, while cortical lesions have no effect. On the other hand, more recent evidence has suggested a more complex set of generators of the FFR. Kuwada et al. (2002), for example, found that deactivating various stations along the auditory neuraxis in rabbits decreased FFR to a certain extent, and that the source of the FFR seemed to vary with frequency, with a more cortical origin for lower frequencies and a more subcortical origin for higher frequencies. This view has been borne out by more recent work; while fMRI, for example, has produced evidence for a modest cortical contribution to the FFR (Coffey, Herholz, Chepesiuk, Baillet, & Zatorre, 2016), research using high-density EEG suggests a rather low upper-frequency limit for cortical contributions to the FFR, with no contribution whatsoever above 150 Hz. Overall, however, the FFR to a complex sound (such as a speech sound) likely mostly reflects processing within subcortical generators. Supporting this view, White-Schwoch, Nicol, Warrier, Abrams

and Kraus (2016) found that trial-by-trial variability measures in the FFR closely tracked timing jitter in the inferior colliculus in guinea pigs, as measured via extracellular recordings.

That the FFR is driven by generators in areas close to the auditory periphery suggests that the influence of attention and cortical state on the response might be attenuated relative to more cortical responses, and this supposition is borne out by experimental data. For example, the FFR is commonly collected when participants are instructed not to attend to the stimulus but instead to watch an unrelated movie (Skoe & Kraus, 2010). Indeed, the FFR can even be recorded while participants sleep (Yamada, Yamane, & Kodera, 1977). Moreover, a number of studies have found no effect of attention on the FFR (see Varghese, Bharadwaj, & Shinn-Cunningham, 2015 for one set of experimental data and a clear summary of prior research on this topic). Similarly, there does not seem to be any effect of phonemic categorization on the FFR (Bidelman, Moreno, & Alain, 2013). However, there is evidence that the FFR can be affected by statistical regularities present in the stimuli. The FFR can, for example, be affected by the novelty of a stimulus relative to its surrounding stimuli (Gao, Zhang, Cheng, Zhou, & Wu, 2014; Slabu, Grimm, & Escera, 2012) as well as its predictability (Lau, Wong, & Chandrasekaran, 2016). Overall, then, the FFR is likely to largely be a measure of implicit auditory processing, although there may be some small effects of explicit (i.e. attention-driven) processing on the lower frequencies of the response (Holmes, Purcell, Carlyon, Gockel, & Johnsrude, 2018).

The FFR is not only of interest to neurophysiologists but has recently been adopted by cognitive neuroscientists and psychologists interested in the neural underpinnings of individual differences in speech perception and language learning. One reason for this is that there are large individual differences in a number of characteristics of the FFR across participants—including its timing, consistency, and strength of spectral encoding—which are highly replicable within participants (Easwar, Scollie, Aiken, & Purcell, 2020; Hornickel, Knowles & Kraus, 2012), with between-session correlations for some metrics reaching $r = 0.88$. Moreover, these metrics have been shown to relate to individual differences in behavioral measures of auditory processing, suggesting that the FFR variability can be a window into variability in auditory skills in the general population. For example, the trial-by-trial consistency of the response is linked to the precision with which individuals are able to synchronize movements to a metronome in both adults (Tierney & Kraus, 2013, 2016) and children (Woodruff Carr, Tierney, White-Schwoch, & Kraus, 2016); the amplitude of the response has been linked to the ability to discriminate stimuli

based on amplitude modulations (Bharadwaj, Masud, Mehraei, Verhulst, & Shinn-Cunningham, 2015); the strength of the representation of the fundamental frequency has been linked to the precision of pitch perception (Carcagno & Plack, 2011; Coffey, Colagrosso, Lehmann, Schonwiesner, & Zatorre, 2016; Krishnan, Bidelman, & Gandour, 2010; Marmel et al., 2013); and the accuracy of vowel formant encoding in the FFR has been linked to vowel recognition performance in participants' L1 (Won et al., 2016).

Given that individual differences in the FFR have been linked to variability in auditory abilities, researchers have also used the FFR to investigate the relationship between language skills and auditory processing, both by comparing language impaired and typically developing populations and by investigating individual differences in L1 skills across the adult population. The proposed mechanism by which auditory processing might impact language acquisition is that variable or imprecise neural representation of acoustic attributes, as reflected by decreased FFR phase-locking, could delay the acquisition of knowledge about phonemic and prosodic categories, which could in turn possibly delay the acquisition of other language skills, including phonological awareness and reading. One consistent finding is that the strength of encoding of the fundamental frequency is linked to the ability to perceive speech in background noise, which has been shown in typically developing children (Anderson, Skoe, Chandrasekaran, Zecker, & Kraus, 2010), young adults (Song, Skoe, Banai, & Kraus, 2011), and older adults (Anderson, Parbery-Clark, Yi, & Kraus, 2011). The FFR has been linked to other language skills as well, particularly reading. For example, Banai et al. (2009) found that phonological awareness and word reading correlated with the strength of encoding of middle harmonics; the participants were a mix of language-impaired and typically developing children, but the analyses related predictors derived from the FFR to continuous variation in outcome measures. Hornickel and Kraus (2013) found that trial-by-trial FFR consistency was linked to reading ability in a large group of children, most of whom were typically developing but a few of whom had been diagnosed as language impaired.

Given that the FFR largely reflects processing in peripheral auditory areas, is mostly unaffected by attention, is linked to individual variation in auditory skills, and has been tied to variation in L1 skills such as speech-in-noise perception and reading, we suggest in the next section that it could be a promising implicit aptitude measure relevant to second language acquisition as well.

Auditory Processing Ability and L2 Learning in Adults

Recently researchers have begun to ask whether the link between auditory processing and successful language acquisition could extend to second language learning. At first, these studies focused on predicting the impact of short-term in-lab training on adults' L2 speech learning. For example, in Lengeris and Hazan (2010), the formant discrimination thresholds of 18 Greek adult learners were related to their training success in English vowel perception and production. After receiving five phonetic training sessions on 14 English vowels over two weeks, learners with lower formant discrimination thresholds before the training tended to perform better at English vowel identification and production tasks after the training. However, the correlation analyses did not take pre-training language skills into account. Thus, it is not clear whether formant discrimination acuity could predict the individual differences in learning during the training sessions or not. Focusing on the suprasegmental aspect of speech perception, Wong and Perrachione (2007) examined auditory processing ability as a predictor of the attainment of non-native Mandarin tone perception during the training. A group of 17 adult American English speakers who reported zero exposure to tone language received training sessions on identifying three mandarin tones (level, rising and falling). According to the regression analysis results, English speakers who could better identify pitch patterns before training showed a higher level of attainment across the training (the initial stage of learning). The authors also examined the effects of musical experience and found that musicians were more likely to be successful learners of Mandarin tones than non-musicians. In this study, the causal effect of auditory processing ability is somewhat clearer, as the link between auditory processing and L2 speech perception learning cannot reflect an influence of language training on auditory skills before the training.

Individual differences in the robustness of the encoding of acoustic information in the FFR have also been shown to predict in-lab non-native speech perception learning. Chandrasekaran, Kraus and Wong (2012) divided a sample of adult English speakers into two groups, one of which had more consistent neural encoding of pitch changes than the other. It was found that the former group demonstrated a faster learning rate across the nine training sessions and almost doubled the latter group's identification ratio of Mandarin tones at the end of training. Findings in these studies suggest that both explicit and implicit auditory processing ability could predict the rate of L2 speech learning in laboratory settings.

Based on these studies, it seems that (a) domain-general auditory processing abilities not only act as an essential foundation for L1 acquisition but could also boost the rate of initial L2 learning; and (b) the initial stage of L2 learning in laboratory settings draws upon both explicit and implicit, pre-conscious auditory processing. However, little research has yet been conducted to investigate the role of auditory processing ability in *naturalistic* L2 immersion contexts, wherein learners acquire a target language through intensive exposure to meaningful, interactive, and authentic input in a similar fashion to L1 acquisition.

In our previous research, we have proposed auditory processing as one component of explicit and implicit aptitude relevant to every stage of naturalistic L2 speech learning (see Saito, Kachlicka, Sun, & Tierney, 2020; Saito, Sun, & Tierney, 2019). Under this framework, explicit auditory processing is defined as one's ability to process sound with some degree of awareness and attention, which we assess via behavioral tasks where participants are aware of the purpose and make careful judgements (e.g., discrimination and reproduction). Implicit auditory processing is defined as early encoding of sound features in subcortical regions of the auditory system, which are relatively unaffected by attention; thus, we assess implicit auditory processing using the FFR.

Thus far, we have cross-sectionally examined the extent to which explicit and implicit auditory processing correlated with L2 English speech perception and production in naturalistic settings among adult learners with different L1 backgrounds (Chinese, Japanese, Spanish, Polish) and varied immersion experience (1–20 years) (e.g., Kachlicka, Saito, & Tierney, 2019; Omote, Jasmin, & Tierney, 2017; Saito et al., 2019, 2020). Omote et al. (2017) studied the link between implicit auditory processing (i.e., FFR) and English speech perception of 25 adult Japanese speakers with varied immersion experience in the UK (M = 2.6 years, SD = 3.1). Participants were tested for the perception of English vowel and consonant contrasts with which this population tends to have difficulty (e.g., /r/-/l/ contrast). According to the results, the more consistent participants' neural responses were from trial to trial, and the more robust the representation of the lower frequencies of sound was, the better they performed in the consonant (but not vowel) perception test. Implicit auditory processing was also found to be linked to L2 speech perception (Kachlicka et al., 2019) and to L2 speech production (Saito et al., 2019, 2020) of L1 Polish and L1 Chinese speakers. As for explicit auditory processing ability, both sound discrimination threshold and rhythmic memory were associated with L2 segmental (Kachlicka et

al., 2019) and prosodic perception (Saito, Sun, Kachlicka, Robert, Nakata, & Tierney, in press). Rhythmic and melodic memory have also been found to relate to the fluency of L2 speech production (Saito et al., 2019, 2020).

Motivation for the Current Study

One limitation of these previous studies mentioned above is that they were cross-sectional in design—auditory processing and language learning success were both measured after a period of immersion. This means that the causality of the link between auditory processing and language learning cannot be established. In particular, it remains possible that enhanced auditory processing is a consequence of successful L2 learning, rather than a predictor (Krizman, Skoe, Marian, & Kraus, 2014; Skoe, Burakiewicz, Figueiredo, & Hardin, 2017). In the current study, we examined the link between explicit and implicit auditory processing and L2 speech learning via a longitudinal design. To answer the research question whether and to what degree explicit and implicit auditory processing ability could predict L2 speech perception gains during the immersion, we assessed, in 46 Chinese learners of English, phonemic and prosodic English speech perception before (Time 1) and after (Time 2) a five-month period of immersion within their first year in the UK, and auditory discrimination, melodic/rhythmic memory and neural encoding of sound before the immersion. Based on the results from previous studies (e.g., Kachlicka et al., 2019), both explicit and implicit aptitude were hypothesized to show predictive power for the L2 speech perception gains.

Method

Participants

A total of 50 Chinese international students were recruited from a few universities in London (majoring in a wide range of subjects including education, engineering, science, linguistics, and management). 46 returned for testing at Time 2 (3 males, 43 females, M_{age} = 23.6, *Range* = 21–29). They were all considered to be late L2 learners of English in the early phase of immersion, as they arrived in the UK after the age of 21 years (*M* = 23, *Range* = 21–

28). At Time 1 they had been in the UK for around 5 months (*M* = 4.5 months, *Range* = 3.9–5.3). Prior to immersion, they had received an extensive amount of foreign language education in China (*M* = 13.5 years, *Range* = 10–19). According to their IELTS scores (above 6.5), their English proficiency levels were intermediate to advanced. Additionally, 11 students reported various lengths of musical training experience (*M* = 6.95 years, *Range* = 0.5–20).

Auditory Processing Measures

Three auditory processing abilities were tested in the current study, including two behavioral measures assessing explicit auditory processing (with conscious awareness) and one neurophysiological measure tapping into implicit processing (without conscious awareness) — (a) sound discrimination threshold, (b) music memory, and (c) neural encoding of sound.

Sound Discrimination Threshold

Following the design in the previous study (Kachlicka et al., 2019), participants' auditory processing threshold was assessed in terms of four different acoustic features of sounds—pitch, formant, duration, and amplitude rise time. During each subtest, participants were asked to listen to a set of three tones (with a half-second interval in between) and identify whether the first or the third one was different from the middle one by pressing the key '1' or '3'.

For each subtest, a total of 100 target stimuli and one baseline stimulus were prepared consisting of artificial sounds varying along continuum of the target acoustic feature (pitch, formant, duration, or amplitude rise time), forming 100 stimulus levels. Higher levels along the continuum were linked to a bigger difference between the baseline and target stimulus and, therefore, easier discrimination. Following the adaptive threshold procedure in Levitt (1971), the tests started with level 50 (i.e., with the target stimulus 50 steps away from the baseline stimulus) and the level changed depending on participants' responses. When the response was incorrect, the difference between stimuli in the next trial became wider (initially by 10 steps) which made the discrimination task easier. When two correct responses were made in a row, the difference between stimuli in the next trial was narrowed (initially by 10 steps), making the task more difficult. Prior to the first incorrect response, however, only a single correct response was

necessary for the task to become more difficult. Once the direction of step changes reversed, the step size of the change became smaller, first to five, then to two, and finally to one step, which was then retained till the end of the test (e.g., 50 → 40 →30 → 35 → 35 → 33 → 33 → 34 → 34 → 33 → 33 → 32 → 32 …). The tests stopped after either 70 trials or eight reversals, and the sound discrimination threshold scores were calculated by averaging the difference levels where the reversals occurred since the third reversal (which is level 34 in the example above). Low threshold scores reflect better sensitivity to differences between sounds.

All stimuli were created using custom MATLAB scripts. Unless described differently, all sounds consisted of 500-ms four-harmonic complex tones with fundamental frequency (F0) at 330Hz and a 15-ms linear ramp at the beginning and end. For the pitch discrimination test, while the baseline stimulus remained at a F0 of 330Hz, that of the target stimuli ranged from 330.3 to 360Hz with a step of 0.3Hz. For the duration discrimination test, the baseline stimulus was 250ms long whereas the target stimuli ranged from 252.5 to 500ms with a step of 2.5ms. For the rise time discrimination test, the baseline stimulus had a rise time of 15ms and that of the target stimuli ranged from 17.85 to 300ms with a step of 2.85ms. For the formant discrimination test, stimuli were complex tones with F0 at 100Hz, the first formant (F1) at 500Hz, the third formant (F3) at 2500Hz and harmonics up to 3000Hz. The second formant (F2) was 1500Hz for the baseline stimulus and 1502–1700Hz with a step of 2Hz for the target stimuli. To form a composite measure of sound discrimination threshold, scores of all four subtests were averaged.

Music Memory

Participants' ability to remember melodic and rhythmic patterns was measured by assessing how accurately they reproduced a melody or rhythm which they listened to three times.

Melodic memory. Ten melodies were prepared as stimuli. Melodies were constructed from a set of five notes, consisting of the first five notes of the major scale, corresponding to frequencies of 220, 246.9, 277.2, 311.1, and 329.6Hz. Each note was 300ms in duration with a 50-ms cosine ramp at the beginning and end of the note. The first note of the melody was always the third pitch. Subsequent notes were then randomly chosen to be either one note higher on the scale or one note lower on the scale. This process repeated until all seven notes were chosen. The

melody could not descend below 220Hz or ascend above 329.6Hz; once the melody reached these limits, the next note was chosen to either be closer to the center of the range or identical to the previous note.

Melodies were repeated three times, with a 1-s interval between each repetition. After each of the melodies was played, five boxes numbered 5–1 were shown on the screen and participants were asked to reproduce the seven notes by clicking one box at a time (starting with Box 3); when each of these boxes was clicked the corresponding note was played. Before the test, participants had a chance to listen to an example and practice with the boxes as much as they needed to get familiar with the five pitches. To calculate response accuracy, a 1-to-1 comparison was made between the notes chosen by the participant and the notes in the target melody, and a percentage score was calculated.

Rhythmic memory. Ten rhythmic patterns (from Povel & Essens, 1985) were prepared as stimuli. The rhythmic patterns consisted of 16 segments, each 200ms, containing either a rest or a drum hit. Nine of the segments contained a drum hit, while the remainder contained a rest. Rhythms were each repeated three times with a 600-ms interval in between. Drum hits consisted of a 150-ms conga drum hit sound acquired from freesound.org. After listening to the stimuli, participants were asked to reproduce the rhythm by pressing the space key. The response time of each pressing was recorded and compared with the drum hit segments in target stimuli. First, the inter-response intervals were quantized by converting them to the nearest multiple of 200 ms. The accuracy of responses was then calculated on a segment-by-segment basis by comparing the content of each segment in the participant's rhythm (i.e. whether it contained a rest or a drum hit) to the corresponding segment in the target rhythm, and a percentage score was calculated. The scores of melodic and rhythmic memory tests were averaged to form a composite measure of music memory.

Neural Encoding of Sound

As discrimination tests require conscious assessment of auditory information, the influence of explicit measures of auditory processing on L2 learning could partially reflect individual differences in attention and memory (Snowling et al., 2018). In contrast, the frequency

following response to sound (FFR; Coffey, Herholz et al., 2016), an electrophysiological response which mirrors the spectro-temporal content of the evoking sound, could be a more pure assessment of auditory processing, as it is relatively unaffected by cognitive and attentional state (Varghese et al., 2015). Thus, the neural encoding of spectral and temporal information of a synthesized speech syllable was examined via the FFR.

Stimulus. A 170-ms consonant-vowel syllable /da/ was synthesized as the stimulus via a Klatt-based synthesizer. It began with a short onset burst of 5ms. Between 5 and 50ms was the transitional stage where the first, second and third formants (F1, F2, F3) changed respectively from 400 to 720Hz, 1700 to 1240Hz, and 2580 to 2500Hz. Then, from 50 to 170ms, these formants remained stable. On the other hand, throughout the stimulus, the fundamental frequency (F0) was constant at 100Hz, while F4, F5 and F6 were constant at 3300 Hz, 3750 Hz, and 4900 Hz, respectively.

Procedure. The /da/ sound was presented repeatedly (6300 times over the course of 20 minutes) at a rate of 4.35Hz, through insert earphones (ER-3; Etymotic Research) at 80dB. To enable separate examination of the amplitude envelope and temporal fine structure of speech (Aiken & Picton, 2008), the stimulus was presented at alternating polarities (i.e. every other stimulus was inverted). To collect electrophysiological responses to the stimulus, a montage of five electrodes was placed on the head of participants—one active electrode on the center of the top of the head (i.e. at Cz), two reference electrodes on the left and right earlobes, and two ground electrodes on the forehead. Contact impedance was maintained beneath 20 kΩ. Continuous electrophysiological data were recorded using a BioSemi ActiveTwo EEG system with a sample rate of 16,384Hz and open filters in ActiView (BioSemi) acquisition software. During the testing session, participants were encouraged to read a book or a magazine of their choice instead of paying attention to the sound. They were also asked to relax their muscles and avoid extraneous body movements.

Data Analyses. All neurophysiological analyses were conducted using custom MATLAB scripts. To begin with, recordings were bandpass filtered between 70 to 2000Hz using a first-order Butterworth filter to isolate the FFR from the cortical evoked response to sound. Then, the

recording was segmented from -30 to 210ms with respect to stimulus presentation. All trials containing amplitude spikes of above 35μV were rejected as artifacts, and then the first 2500 artifact-free responses to each stimulus polarity (5000 total sweeps) were selected for further analysis.

The accuracy of neural sound encoding was measured via inter-trial phase-locking analysis. This analysis reveals the degree of temporal consistency in the response across trials at each frequency. Our use of inter-trial phase locking analysis rather than spectral analysis of the average response was motivated by prior work showing that inter-trial phase-locking demonstrates a comparatively greater signal-to-noise ratio (Zhu et al., 2013). For each trial, a Hanning-windowed fast Fourier transform was conducted over a response time window between 10 and 180ms (10–180ms for F0 encoding, 60–180ms for F1 & F2 encoding). The outcome of this procedure consists of a complex vector for each trial with information about the amplitude and phase of the neural response. Next, these vectors were converted to unit vectors which retained only the phase information and were averaged. Greater length of the averaged vector indicates similar phases across the unit vectors. Thus, the length was taken as the measure of inter-trial phase consistency, which varies from 0 (no consistency/phases uniformly distributed) to 1 (perfect consistency/phases identical across trials). It is worth noting that there was an extra step for the analysis of F1 & F2 encoding before the inter-trial phase-locking procedure—the phases of trials corresponding to one polarity were shifted 180 degrees to emphasize the temporal fine structure of the stimulus, which enables analysis of neural encoding of the higher-frequency formants (Aiken & Picton, 2008).

In line with our previous study (Kachlicka et al. 2019), we focused on three frequencies that were particularly important in the evoking sound: 100Hz (F0), 720Hz (F1), and 1240Hz (F2). Neural encoding of F0 was calculated as the maximum inter-trial phase coherence between 80 and 120Hz, whereas neural encoding of F1 was calculated as the maximum inter-trial phase coherence between 680 and 720Hz, and neural encoding of F2 was calculated as the average of the average of the maximum inter-trial phase coherence between 1180 and 1220Hz and the maximum between 1280 and 1320 Hz. To obtain a composite score of neural encoding of sound, an average score of the phase-locking consistency at F0, F1 and F2 was calculated for each participant.

L2 and Musical Experience Measure

Participants reported their experience of L2 interaction and musical training via an online questionnaire. Although participants' length of residence in the UK was similar, daily use of the target language varied widely across participants. In the current study, L2 experience was measured by recording participants' interactive L2 use, which could be crucial to L2 speech learning (e.g., Moyer, 2011). A survey was conducted at Time 2 where students reported retrospectively the weekly hours spent on L2 speaking in professional, home and social settings during the 5-month immersion. The hours were added up to reflect the amount of L2 experience. As for musical experience, 11 participants had received regular formal training by Time 1. In the questionnaire, they provided information about the length of training in years and the focus of training. Due to the small number of participants with musical training experience, the data was encoded categorically (0 = no experience at all, 1 = any experience).

L2 Proficiency Measures

To examine the degree of improvement in phonological knowledge of the L2, participants' ability to differentiate English speech contrasts at the segmental and suprasegmental level (i.e., speech perception) was assessed before and after the immersion period. Following the previous study (Kachlicka et al., 2019), participants were asked to listen to a word or sequence of words and choose the word or phrase which best matched what they heard from two options shown on the screen by pressing the keys '1' (left) or '2' (right). The stimuli were minimal pairs comprising vowel contrasts (e.g., /æ/ vs. /e/), consonant voicing contrasts (e.g., /d/ vs. /t/) and phrases which differed in contrastive focus (i.e., READ books versus read BOOKS). There were 20 pairs for each contrast. All stimuli were produced by a native speaker of Southern British English. The test was run in MATLAB. The speech perception scores were calculated as the percentage of correct answers out of the 20 trials. Participants' performance on consonant perception at Time 1 was largely at ceiling; only two participants did not achieve a perfect score, and even these participants answered only a single item incorrectly. As a consequence, only data from the vowel and prosody items was analyzed further.

Contrastive focus stimuli were taken from the Multidimensional Battery of Prosody Perception (MBOPP; Jasmin, Dick, & Tierney, 2020). This test battery consists of minimal pairs of recorded phrases which are identical lexically but differ on a single prosodic feature. The speech morphing software STRAIGHT (Kawahara & Irino, 2005) was used to morph these two phrases onto one another, so that they could be set to differ only in their durational and pitch properties. The duration and pitch cues to the location of contrastive focus were then set to 60% of their original size, in an attempt to avoid ceiling effects.

Procedure

Data was collected in a lab at the Department of Psychological Sciences at Birkbeck, University of London. All auditory processing and speech perception tasks were conducted at Time 1; both speech perception tasks were also conducted at Time 2 using the same materials, together with the EEG test and the survey for L2 and musical experience, but data from the EEG test at Time 2 is not analyzed here. Tasks were administered in the following order: Sound Discrimination Test, Speech Perception Test, Music Memory Test, and Experience Survey. Finally, the FFR was recorded. All instructions were delivered in both English and Chinese to avoid any misunderstandings of the procedure. The testing sessions lasted for approximately 2 hours at Time 1 and for around 1.5 hour at Time 2.

Reliability Analyses

The test-retest reliability of all measures was examined by correlating performance at Time 1 and Time 2 for each measure. The reliability of all three auditory processing measures ranged from .70 to 0.86, which can be taken as acceptable (Lance, Butts, & Michels, 2006). FFR phase-locking was calculated based on the 45 participants who completed the EEG test at both Time 1 and Time 2 ($r = 0.83$, $p < .001$). The sound discrimination task and the music memory task were not conducted at Time 2 in this study. Thus, we conducted a separate project for the test-retest reliability, where we recruited and asked a total of 30 L1 and L2 English users to take the same sound discrimination task and music memory task twice in two consecutive days. According to the correlation analyses, their test-retest performance demonstrated relatively

strong associations—$r = .70$, $p < .001$ for sound discrimination threshold, and $r = .86$, $p < .001$ for music memory (for details, see Supplementary Material; see also our full report in Saito, Sun, & Tierney, 2020). As for the speech perception measures, although the reliability of the prosody perception test was acceptable ($r = 0.68$, $p < .001$), the reliability of the vowel perception test was low ($r = 0.47$, $p < .001$).

A total of three independent variables (sound discrimination threshold, music memory, neural encoding of sound) and two dependent variables (L2 vowel and prosody perception scores at Time 2) were entered into the analysis. First, a set of paired-samples t-tests was run on the linguistic measures to show if participants made any significant improvement in L2 knowledge from Time 1 to Time 2. For those measures that demonstrated significant gains, the Time 2 scores were related to independent variables via partial correlation analyses (with Time 1 scores controlled for) to reveal any predictors of L2 gains. Subsequently, multiple regression analyses were conducted, with the auditory processing measures as predictors and Time 2 score as the outcome variable, and with Time 1 scores, L2 and musical experience controlled for.

According to Shapiro-Wilk's test, scores of music memory, neural encoding of sound, L2 experience and prosody perception at Time 1 were not normally distributed ($p < .05$). Thus, non-parametric Spearman correlations were conducted. As for the multiple regression analyses, the residuals were normally distributed.

Results

Gains in L2 Speech Perception

To investigate whether auditory processing ability can predict the longitudinal development of L2 speech perception, we first examined whether and to what degree participants improved in vowel and prosody perception tasks from Time 1 to Time 2. Given that some participants' performance at Time 1 already reached ceiling (i.e. 100% correct performance) and had no room for improvement, their data was excluded from the statistical analyses, which left $N = 44$ for the vowel perception test and $N = 31$ for the prosody perception test.

As gain scores (Time 2 - Time 1) of both vowel and prosody perception (based on the downsized datasets) were normally distributed according to Shapiro-Wilk's test ($p > 0.1$), the

Time 1 and Time 2 scores of both speech perception dimensions were submitted to paired-samples t-tests. The results (summarized in Table 1) showed significant gains in prosody perception scores over time, $t(30) = 3.22$, $p = .003$, but not in vowel perception scores, $t(43) = -0.27$, $p = .788$. Therefore, the subsequent analyses only focused on predictors of L2 prosody perception gain scores based on the $N = 31$ dataset. According to Plonsky and Oswald's (2014) field-specific benchmarks ($d = .60$ as small, 1.00 as medium, 1.40 as large), the improvement participants made in prosody perception showed a small effect size ($d = .43$).

Table 1

Results of Paired-samples T-tests assessing L2 Speech Perception (Time 1 and Time 2)

	Mean Time 1 (SD1)	*Mean Time 2 (SD2)*	*95% CI of the difference*	*t-value*	*df*	*p-value*
Vowel perception	0.836 (0.090)	0.832 (0.126)	-0.04, 0.03	0.27	43	$p = .788$
Prosody perception	0.819 (0.126)	0.873 (0.123)	0.02, 0.09	3.22	30	$p = .003$

Note. The scores of L2 speech perception were calculated as the percentages of correct answers.

Auditory Processing Ability Profiles

The descriptive results of the three auditory processing ability measures are summarized in Table 2. Participants showed individual variability to various degrees in terms of their auditory processing abilities at Time 1. In order to investigate the interrelationships between the three independent variables, a set of Spearman's non-parametric correlation analyses was conducted. To adjust for multiple comparisons, the alpha level was set at .017 via the Bonferroni correction. As shown in Table 3, no significant correlation was found between the three auditory processing measures. More specifically, there was no evidence that (a) the neural encoding of sound, which was assumed to tap into the implicit dimension of auditory processing, was related to the explicit auditory processing measures; and that (b) the two explicit auditory processing measures, sound discrimination threshold and music memory, were associated with each other.

Table 2

Descriptive Statistics of Participants' Auditory Processing Ability Profiles

	M	*SD*	*Range*		*95% CI*	
			Min	*Max*	*Lower*	*Upper*
Sound discrimination threshold (1–100)	22.82	8.12	7.86	42.53	19.85	25.80
Music memory (0–1)	0.70	0.14	0.47	0.93	0.65	0.75
Neural encoding of sound (0–1)	0.12	0.04	0.05	0.22	0.11	0.13

Note. For sound discrimination threshold, lower scores indicate better performance.

Table 3

Correlations among Auditory Processing Ability Variables

	Music memory		Neural encoding of sound	
	r	*p*	*r*	*p*
Sound discrimination threshold	-.283	.123	-.015	.938
Music memory			.163	.380

Note. No correlation reached the significance level at $p < .017$ (Bonferroni corrected).

Auditory Processing and Gains in L2 Prosody Perception

In order to examine which of the three independent variables (i.e., sound discrimination threshold, music memory, and neural encoding of sound) could predict the gains in L2 prosody perception from Time 1 to Time 2, we conducted a set of non-parametric partial Spearman's correlation analyses. Auditory processing abilities were submitted to the analyses as predictors, with the prosody perception scores at Time 2 as the dependent variable and Time 1 scores controlled for as a covariate. To adjust for multiple comparisons, the alpha level was set at .017 via the Bonferroni correction.

As summarized in Table 4, L2 prosody perception scores at Time 2 showed a significant and positive correlation with music memory ($r = .456$, $p = .011$), which suggests that music

memory was a predictor of L2 prosody perception gains. See Figure 1 for scatterplots displaying the relationship between music memory and L2 prosody perception scores at Time 2. According to Plonsky and Oswald's (2014) field-specific benchmarks (r = .25 as small, .40 as medium, .60 as large), the strength of the correlations indicated that the role of music memory in L2 speech perception development could be considered as "medium". On the other hand, implicit auditory processing ability (i.e., neural encoding of sound) and explicit auditory discrimination thresholds did not show correlations with gains in L2 prosody perception.

Table 4

Results of Partial Correlation Analyses of Auditory Processing and L2 Prosody Perception at Time 2

	L2 prosody perception at Time 2	
	r	*p*
Sound discrimination threshold	-.290	.120
Music memory	.456	.011*
Neural encoding of sound	.014	.942

Note. *p < .017 (Bonferroni corrected). Time 1 scores of L2 prosody perception were controlled for.

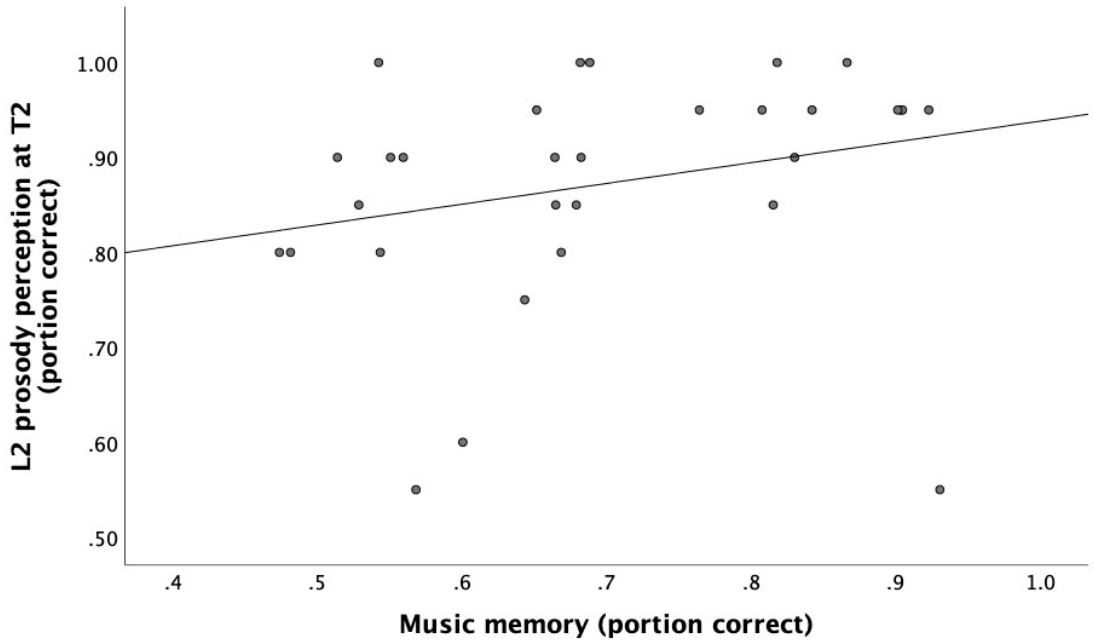

Figure 1. Scatterplot displaying the relationship between music memory and L2 prosody perception scores at Time 2

Predictors of Gains in L2 Prosody Perception

To determine the predictors of the gains in L2 prosody perception and the amount of variance they could explain, music memory, Time 1 prosody perception scores, and degree of L2 experience (i.e., weekly hours of L2 speaking) were submitted to stepwise multiple regression analyses as independent variables. Sound discrimination threshold and neural encoding of sound did not enter into the model because they were not correlated with Time 2 prosody perception scores after covarying for Time 1 scores. As participants with musical experience showed better music memory than those without, $t(29) = 4.66$, $p < .001$, to test if the correlation between music memory and L2 prosody perception gains was mediated by musical experience, the musical experience measure (0 = no experience at all, 1 = any experience) was also submitted to stepwise multiple regression analyses as an independent variable. Prosody perception score at Time 2 was submitted as a dependent variable.

As shown in Table 5, only music memory remained in the model as a predictor, accounting for 10.5% of the variance in L2 prosody perception at Time 2, apart from the L2

prosody perception Time 1 scores (explaining 52.6% of the variance), whereas the amount of L2 interaction and musical training experience were excluded from the model. According to the interpretations of effect sizes by Plonsky and Oswald (2014), the variance explained by music memory corresponded to a small-to-medium effect size (6.25%< R^2 <16%).

Table 5

Significant Results of Multiple Regression Analyses on Auditory Processing as Predictors of L2 Prosody Perception at Time 2

Variable	B	95% CI for B		SE B	β	R^2	$\triangle R^2$
		Lower	Upper				
Step 1						.526	.526***
Constant	.290**	.078	.502	.104			
Time 1 scores of L2 Prosody Perception	.711***	.455	.967	.125	.726***		
Step 2						.632	.105**
Constant	.073	-.175	.320	.121			
Time 1 scores of L2 Prosody Perception	.740***	.509	.971	.113	.755***		
Music memory	.277**	.076	.478	.098	.326**		

Note. **p < .01, ***p < .001.

Discussion

The current study longitudinally examined the link between domain-general auditory processing (explicit, implicit) and L2 speech learning during the early stage of immersion with a pretest-posttest design. A total of 46 Chinese learners of English were tested on their L2 segmental (vowel and consonant) and prosodic (contrastive focus) speech perception at the beginning and the end of a five-month immersion in an English-speaking environment (i.e., the UK). While the 31 participants who performed below ceiling in prosody perception at the beginning of immersion demonstrated significant improvement in prosody perception after immersion, no significant improvement in L2 segmental perception was found at the group level.

The learning difficulty of vowels is in line with Munro and Derwing (2008), who found that Mandarin speakers' performance on English vowel production generally stabilized after a half year of immersion. However, our finding of significant gains in L2 prosody but not vowel perception could also just reflect the relative reliability of these two measures, given that the vowel perception test showed particularly low reliability ($r = 0.47$).

Based on our previous research (Kachlicka et al., 2019; Saito et al., 2019), three types of auditory processing abilities were assessed at Time 1. Behavioral tests assessing relatively explicit auditory processing included measurements of sensitivity to sounds differing in spectral and temporal features (i.e., sound discrimination threshold) and tests of the ability to remember and reproduce melodic and rhythmic patterns of non-verbal audio input (i.e., music memory). Implicit pre-conscious auditory sensitivity was measured via the frequency following response to sound, an electrophysiological response that reproduces the spectro-temporal characteristics of the evoking stimulus.

According to the correlation analyses, these measures evaluated three independent domains of auditory processing abilities. Among them, the results of multiple regression analysis indicated that music memory was the sole predictor of L2 prosody perception gains, accounting for 10.5% of the variance, even when music training experience was entered as a potential predictor. Implicit auditory processing (neural encoding of sound) and explicit auditory discrimination thresholds, however, did not explain L2 prosody perception gains from a short immersion. It is noteworthy that according to the multiple regression analyses, the amount of authentic and interactive L2 input during immersion did not relate to the extent of L2 speech perception gains. This may reflect the relatively short duration of the L2 exposure here, which may have been too brief for L2 input characteristics to have an effect on speech perception gains. Our findings provide more support to the view that the outcomes of post-pubertal L2 speech learning can be influenced not only by experience-related factors but also by individual differences in learners' perceptual-cognitive abilities (Saito, 2019; Saito et al., 2019, 2020), by tapping into domain-general perceptual abilities. In what follows, we discuss how individual differences in L2 prosody acquisition could reflect different types of explicit and implicit auditory processing—memory, discrimination, and neural encoding.

Explicit Auditory Processing and L2 Speech Acquisition

The longitudinal relationship between music memory and L2 prosody perception gains found in this study extends the findings of previous cross-sectional studies which reported that learners who performed better on L2 speech perception and production tend to have stronger music skills (Kachlicka et al., 2019; Saito et al., 2019, 2020). Here, we show that music skills (rhythmic and melodic memory) assessed at the very beginning of a period of immersion can predict subsequent L2 speech learning gains, which suggests that the relationship between auditory processing and L2 language learning does not merely reflect the effects of successful language learning, but instead that auditory processing abilities may play a causal role in helping determine the speed of L2 acquisition (cf. Snowling et al., 2018). In other words, learners who can better detect and reproduce the timing and pitch contour of sounds may find it easier to extract pitch and duration information in order to identify contrastive focus (and, potentially, other prosodic features such as phrase boundaries).

However, we agree that even when auditory processing at an earlier time point predicts language learning success at a later time point, it remains possible that this association is driven by a third factor related to auditory processing, such as socioeconomic status or modality-general cognitive skills (e.g., working memory or attention) (Doughty, 2019; Linck et al., 2013). As a result, the causality of the link between auditory processing and language learning remains an important topic for future research (Mueller, Friederici, & Mañnel, 2012), which could be addressed via intervention studies (e.g., Li & DeKeyser, 2017).

We found no correlation between L2 prosody perception gains and sound discrimination thresholds, conflicting with our earlier findings of a link between auditory discrimination and L2 speech perception (Kachlicka et al., 2019). This could reflect the different stages of immersion in these two studies (experienced vs. inexperienced), or the different L1s spoken by the participants (Polish vs. Mandarin). In our recent investigations with adult L2 learners with various lengths of immersion and L1 backgrounds (Saito, Sun et al., in press), there is some preliminary evidence that learners' ability to detect acoustic details of sound (measured via sound discrimination tasks) can predict the extent to which L2 learners can continue to improve and attain advanced L2 proficiency, provided an ample amount of L2 immersion experience though a longer period of immersion. It is probable that such perceptual acuity plays a crucial role especially in the mid-to-

ultimate phases of L2 learning (see Linck et al., 2013), while learners' ability to remember melodic and rhythmic patterns (measured via reproduction tasks) appears to be equally linked to various stages and contexts of L2 learning and attainment (Saito, Tran, Suzukida, & Tierney, in press).

Implicit Auditory Processing and L2 Speech Acquisition

Importantly, a relationship was not identified between implicit auditory processing ability (as assessed via neural encoding of speech) and L2 prosody perception learning, again conflicting with Kachlicka et al. (2019) and Omote et al. (2017), who found a robust relationship between FFR encoding and speech perception after several years of L2 English immersion. There are several possible explanations for this null finding. First, the amount of immersion (five months) participants had in this study might not be enough for implicit auditory processing ability to play a role. After receiving over ten years of formal L2 instruction in China prior to their arrival in the UK, learners seemed to rely on explicit processing of the target language within the first year of immersion. Thus, the effects of implicit auditory processing on L2 speech learning may become more evident when participants have accumulated enough immersion at a later phase of naturalistic learning (Granena, 2013; Suzuki & DeKeyser, 2015). Indeed, our work has shown that implicit auditory processing could explain variances among L2 learners with sufficiently long length of residence in L2 speaking countries (e.g., 5-10 years; Kachlicka et al., 2019; Saito et al., 2020).

An alternate perspective is that the focus of Kachlicka et al. (2019) and Omote et al. (2017) was on L2 segmental perception while the focus of this study was on suprasegmental perception; thus, the inconsistent findings may reflect the different roles of implicit auditory processing in segmental vs. suprasegmental speech learning. According to the results of Saito et al. (2019), while segmental L2 speech production was related to both explicit and implicit auditory processing, suprasegmental production was related only to rhythmic memory. There seems to be a possibility that while segmental learning may draw upon the precision of encoding of auditory dimensions, as reflected in the robustness of the FFR, suprasegmental learning may instead draw upon the ability to remember changes in rhythmic and melodic (i.e. durational and

pitch) patterns. However, more investigation on both segmental and suprasegmental learning are needed to show a clearer picture.

A third explanation concerns the extent to which the outcome measures in this study (vowel and prosody perception tasks) required L2 speakers to rely on their implicit language aptitude. Compared to producing spontaneous speech in the target language, differentiating contrasts of words or short phrases is much easier to monitor explicitly and thus likely to relate to explicit aptitude rather than implicit aptitude (cf. Skehan, 2016). Finally, in this small dataset ($N = 31$), participants' variability in L2 speech learning was limited (e.g., the ceiling effects found in L2 speech perception measures), which may confound the predictive power of certain auditory processing measures.

Overall, the results from the current study and previous studies suggest different roles for explicit and implicit auditory processing abilities in L2 speech learning. On the one hand, implicit auditory processing seems to have a more salient effect on the ultimate attainment of certain aspects (arguably those more difficult to be mastered) of L2 speech learning than on the initial learning rate. On the other hand, explicit auditory processing could contribute to various stages and aspects of L2 speech learning.

Conclusions and Future Directions

The current study is a preliminary longitudinal investigation of the effects of auditory processing ability on L2 speech learning with a pretest-posttest design. Focusing on the L2 speech perception gains from a short phase of early immersion, the results support a predictive role for explicit but not implicit auditory processing in driving gains in prosody perception. Here we acknowledge several methodological limitations and call for more future studies to investigate the impact of auditory processing among a larger number of participants with more balanced gender distributions, as well as more varied L1 backgrounds (tonal vs. non-tonal), language learning experience (classroom vs. immersion), and proficiency levels (cf. Saito, Sun et al., in press).

Depending on their L1s (e.g., tonal vs. non-tonal), L2 learners may recruit different spectro-temporal cues to extract information from the auditory input (Jasmin, Sun, & Tierney, 2021). Therefore, it would be interesting to compare the impact of spectral and temporal

perception abilities on L2 acquisition among learners with various L1-L2 pairings. Moving forward, future studies could also examine the longitudinal development of various L2 skills and the dynamic interactions between different types of auditory processing, experience and L2 performance over a longer period of immersion. More measures of auditory processing ability tapping into explicit and implicit dimensions should also be encouraged and their relationships should be explored. Although we argue that sound discrimination and music memory tasks draw heavily on explicit processing and FFR on implicit processing, we cannot conclusively rule out the possibility that both explicit and implicit processes contribute to participants' performance in all of these tasks. Developing a wider battery of measures of auditory processing would also enable researchers to begin to gain a clearer picture of the ways in which auditory processing can be fractionated into different skills, and the relative importance of these skills for L2 learning. More reliable measures of skills for L2 learning are also worth exploring, to avoid the lack of variability or gains over time caused by low reliability (such as the vowel perception task in the current study). Finally, more research is needed to test our tentative hypothesis that more precise auditory processing leads to more successful L2 speech learning. One intriguing direction is to further investigate the causal relationship between audition and acquisition by conducting random-assignment intervention studies with control groups to examine whether and to what degree auditory training at the outset of L2 immersion can enhance L2 speech acquisition.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in second language acquisition, 30*(4), 481-509.

Ahissar, M., Protopapas, A., Reid, M., & Merzenich, M. M. (2000). Auditory processing parallels reading abilities in adults. *Proceedings of the National Academy of Sciences, 97*(12), 6832-6837.

Aiken, S. J., & Picton, T. W. (2008). Envelope and spectral frequency-following responses to vowel sounds. *Hearing research, 245*, 35-47.

Anderson, S., Parbery-Clark, A., Yi, H. G., & Kraus, N. (2011). A neural basis of speech-in-noise perception in older adults. *Ear and hearing, 32*(6), 750.

Anderson, S., Skoe, E., Chandrasekaran, B., Zecker, S., & Kraus, N. (2010). Brainstem correlates of speech-in-noise perception in children. *Hearing research, 270*(1-2), 151-157.

Banai, K., Hornickel, J., Skoe, E., Nicol, T., Zecker, S., & Kraus, N. (2009). Reading and subcortical auditory function. *Cerebral cortex, 19*(11), 2699-2707.

Banai, K., Nicol, T., Zecker, S. G., & Kraus, N. (2005). Brainstem timing: implications for cortical processing and literacy. *Journal of Neuroscience, 25*(43), 9850-9857.

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015). Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience, 35*(5), 2161-2172.

Bialystok, E., & Fröhlich, M. (1978). Variables of classroom achievement in second language learning. *The Modern Language Journal, 62*(7), 327-336.

Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage, 79*, 201-212.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and cognitive processes, 25*(7-9), 1044-1098.

Carcagno, S., & Plack, C. J. (2011). Subcortical plasticity following perceptual learning in a pitch discrimination task. *Journal of the Association for Research in Otolaryngology, 12*(1), 89-100.

Carroll, J. B., & Sapon, S. (1959) *Modern Language Aptitude Test: Form A*. New York: Psychological Corporation.

Casini, L., Pech-Georgel, C., & Ziegler, J. C. (2018). It's about time: revisiting temporal processing deficits in dyslexia. *Developmental Science, 21*(2), e12530.

Chandrasekaran, B., Kraus, N., & Wong, P. C. (2012). Human inferior colliculus activity relates to individual differences in spoken language learning. *Journal of neurophysiology, 107*(5), 1325-1336.

Coffey, E. B., Colagrosso, E. M., Lehmann, A., Schönwiesner, M., & Zatorre, R. J. (2016). Individual differences in the frequency-following response: relation to pitch perception. *PLoS One, 11*(3).

Coffey, E. B., Herholz, S. C., Chepesiuk, A. M., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature communications, 7*(1), 1-11.

Curtin, C., Avner, A., & Smith, L. A. (1983). The Pimsleur Battery as a predictor of student performance. *Modern Language Journal, 67*(1), 33-40.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of memory and language, 31*(2), 218-236.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in second language acquisition, 22*(4), 499-533.

Doughty, C. J. (2019). Cognitive language aptitude. *Language learning, 69*, 101-126.

Easwar, V., Scollie, S., Aiken, S., & Purcell, D. (2020). Test-Retest Variability in the Characteristics of Envelope Following Responses Evoked by Speech Stimuli. *Ear and Hearing*, *41*(1), 150-164.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *The Journal of the Acoustical Society of America, 97*(3), 1893-1904.

Flaugnacco, E., Lopez, L., Terribili, C., Zoia, S., Buda, S., Tilli, S., ... & Schön, D. (2014). Rhythm perception and production predict reading abilities in developmental dyslexia. *Frontiers in human neuroscience, 8*, 392.

Galbraith, G. C., Bhuta, S. M., Choate, A. K., Kitahara, J. M., & Mullen Jr, T. A. (1998). Brain stem frequency-following response to dichotic vowels during attention. *Neuroreport, 9*(8), 1889-1893.

Gao, P. P., Zhang, J. W., Cheng, J. S., Zhou, I. Y., & Wu, E. X. (2014). The inferior colliculus is involved in deviant sound detection as revealed by BOLD fMRI. *Neuroimage, 91*, 220-227.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences, 99*(16), 10911-10916.

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63*(4), 665-703.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311-343.

Grube, M., Kumar, S., Cooper, F. E., Turton, S., & Griffiths, T. D. (2012). Auditory sequence analysis and phonological skill. *Proceedings of the Royal Society B: Biological Sciences, 279*(1746), 4496-4504.

Holmes, E., Purcell, D. W., Carlyon, R. P., Gockel, H. E., & Johnsrude, I. S. (2018). Attentional modulation of envelope-following responses at lower (93–109 Hz) but not higher (217–233 Hz) modulation rates. *Journal of the Association for Research in Otolaryngology, 19*(1), 83-97.

Hornickel, J., Knowles, E., & Kraus, N. (2012). Test-retest consistency of speech-evoked auditory brainstem responses in typically-developing children. *Hearing research*, *284*(1-2), 52-58.

Hornickel, J., & Kraus, N. (2013). Unstable representation of sound: a biological marker of dyslexia. *Journal of Neuroscience, 33*(8), 3500-3504.

Jaffe-Dax, S., Kimel, E., & Ahissar, M. (2018). Shorter cortical adaptation in dyslexia is broadly distributed in the superior temporal lobe and includes the primary auditory cortex. *ELife, 7*, e30018.

Jasmin, K., Dick, F., & Tierney, A. T. (2020). The Multidimensional Battery of Prosody Perception (MBOPP). *Wellcome Open Research, 5*(4), 4.

Jasmin, K., Sun, H. & Tierney, A. T. (2021). Effects of language experience on domain-general perceptual strategies. *Cognition*, *206*, 104481. https://doi.org/10.1016/j.cognition.2020.104481

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and language, 192*, 15-24.

Kawahara, H., & Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In P. Divenyi (Ed.), *Speech separation by humans and machines* (pp. 167-180). Boston, MA: Springer.

Kiren, T., Aoyagi, M., Furuse, H., & Koike, Y. (1994). An experimental study on the generator of amplitude-modulation following response. *Acta oto-laryngologica. Supplementum, 511*, 28-33.

Krishnan, A., Bidelman, G. M., & Gandour, J. T. (2010). Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. *Hearing research, 268*(1-2), 60-66.

Krizman, J., Skoe, E., Marian, V., & Kraus, N. (2014). Bilingualism increases neural response consistency and attentional control: Evidence for sensory and cognitive coupling. *Brain and language, 128*(1), 34-40.

Kuwada, S., Anderson, J. S., Batra, R., Fitzpatrick, D. C., Teissier, N., & D'Angelo, W. R. (2002). Sources of the scalp-recorded amplitude-modulation following response. *Journal of the American Academy of Audiology, 13*(4), 188-204.

Lau, J. C., Wong, P. C., & Chandrasekaran, B. (2017). Context-dependent plasticity in the subcortical encoding of linguistic pitch patterns. *Journal of neurophysiology, 117*(2), 594-603.

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America, 128*(6), 3757-3768.

Levitt, H. C. C. H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America, 49*(2B), 467-477.

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.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics, 36*(3), 385-408.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition, 38*(4), 801-842.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., ... & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language learning, 63*(3), 530-566.

Marmel, F., Linley, D., Carlyon, R. P., Gockel, H. E., Hopkins, K., & Plack, C. J. (2013). Subcortical neural synchrony and absolute thresholds predict frequency discrimination independently. *Journal of the Association for Research in Otolaryngology, 14*(5), 757-766.

Marslen-Wilson, W. D., Tyler, L. K., Warren, P., Grenier, P., & Lee, C. S. (1992). Prosodic effects in minimal attachment. *The Quarterly Journal of experimental psychology, 45*(1), 73-87.

McArthur, G. M., & Bishop, D. V. (2005). Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study. *Brain and language, 94*(3), 260-273.

Meara, P. (2005). *LLAMA language aptitude tests: The manual.* Swansea: Lognostics.

Moushegian, G., Rupert, A. L., & Stillman, R. D. (1973). Scalp-recorded early responses in man to frequencies in the speech range. *Electroencephalography and clinical neurophysiology, 35*(6), 665-667.

Moyer, A. (2011). An investigation of experience in L2 phonology: Does quality matter more than quantity? *Canadian Modern Language Review, 67*(2), 191-216.

Mueller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences, 109*(39), 15953-15958.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language learning, 58*(3), 479-502.

Omote, A., Jasmin, K., & Tierney, A. (2017). Successful non-native speech perception is linked to frequency following response phase consistency. *cortex, 93*, 146-154.

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (form S).* Harcourt, Brace and world, Incorporated.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

Povel, D. J., & Essens, P. (1985). Perception of temporal patterns. *Music Perception: An Interdisciplinary Journal, 2*(4), 411-440.

Saito, K. (2019). The role of aptitude in second language segmental learning: The case of Japanese learners' English /r/ pronunciation attainment in classroom settings. *Applied Psycholinguistics, 40*, 183-204.

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language, 115*, 104168.

Saito, K., Sun, H., & Tierney, A. (2019). Explicit and implicit aptitude effects on second language speech learning: Scrutinizing segmental and suprasegmental sensitivity and performance via behavioural and neurophysiological measures. *Bilingualism: Language and Cognition, 22*(5), 1123-1140.

Saito, K., Sun, H., & Tierney, A. (2020). Brief report: Test-retest reliability of explicit auditory processing measures. *bioRxiv*.

Saito, K., Sun, H., Kachlicka, M., Robert, J., Nakata, T., & Tierney, A. (in press). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*.

Saito, K., Tran, M., Suzukida, Y., & Tierney, A. (in press). Auditory processing partially explains L2 speech learning in classroom settings: A review and generalization study. *Language Learning*.

Skehan, P. (2002) Theorizing and updating aptitude. In P. Robinson (ed.), *Individual Differences and Instructed Language Learning* (pp. 69-93). Amsterdam: Benjamins.

Skehan, P. (2016). Tasks versus conditions: Two perspectives on task research and their implications for pedagogy. *Annual Review of Applied Linguistics, 36*, 34-49.

Skoe, E., Burakiewicz, E., Figueiredo, M., & Hardin, M. (2017). Basic neural processing of sound in adults is influenced by bilingual experience. *Neuroscience, 349*, 278-290.

Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and hearing*, *31*(3), 302.

Slabu, L., Grimm, S., & Escera, C. (2012). Novelty detection in the human auditory brainstem. *Journal of Neuroscience, 32*(4), 1447-1452.

Smith, J. C., Marsh, J. T., & Brown, W. S. (1975). Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. *Electroencephalography and clinical neurophysiology, 39*(5), 465-472.

Snowling, M. J., Gooch, D., McArthur, G., & Hulme, C. (2018). Language skills, but not frequency discrimination, predict reading skills in children at risk of dyslexia. *Psychological science, 29*(8), 1270-1282.

Sohmer, H., Pratt, H., & Kinarti, R. (1977). Sources of frequency following responses (FFR) in man. *Electroencephalography and clinical neurophysiology, 42*(5), 656-664.

Song, J. H., Skoe, E., Banai, K., & Kraus, N. (2011). Perception of speech in noise: neural correlates. *Journal of cognitive neuroscience, 23*(9), 2268-2279.

Sparks, R. L., Artzer, M., Ganschow, L., Siebenhar, D., Plageman, M., & Patton, J. (1998). Differences in native-language skills, foreign-language aptitude, and foreign-language grades among high-, average-, and low-proficiency foreign-language learners: Two studies. *Language testing, 15*(2), 181-216.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*(4), 860-895.

Tierney, A., & Kraus, N. (2013). The ability to move to a beat is linked to the consistency of neural responses to sound. *Journal of Neuroscience, 33*(38), 14981-14988.

Tierney, A., & Kraus, N. (2016). Getting back on the beat: links between auditory–motor integration and precise auditory processing at fast time scales. *European Journal of Neuroscience, 43*(6), 782-791.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science, 34*(3), 434-464.

Varghese, L., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2015). Evidence against attentional state modulating scalp-recorded auditory brainstem steady-state responses. *Brain Research, 1626*, 146-164.

Walker, M. M., Shinn, J. B., Cranford, J. L., Givens, G. D., & Holbert, D. (2002). Auditory temporal processing performance of young adults with reading disorders. *Journal of Speech, Language, and Hearing Research, 45*(3), 598-605.

White-Schwoch, T., Nicol, T., Warrier, C. M., Abrams, D. A., & Kraus, N. (2017). Individual differences in human auditory processing: insights from single-trial auditory midbrain activity in an animal model. *Cerebral Cortex, 27*(11), 5095-5115.

Won, J. H., Tremblay, K., Clinard, C. G., Wright, R. A., Sagi, E., & Svirsky, M. (2016). The neural encoding of formant frequencies contributing to vowel identification in normal-hearing listeners. *The Journal of the Acoustical Society of America, 139*(1), 1-11.

Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*(4), 565-585.

Woodruff Carr, K., Tierney, A., White-Schwoch, T., & Kraus, N. (2016). Intertrial auditory neural stability supports beat synchronization in preschoolers. *Developmental Cognitive Neuroscience, 17*, 76-82.

Yamada, O., Yamane, H., & Kodera, K. (1977). Simultaneous recordings of the brain stem response and the frequency-following response to low-frequency tone. *Electroencephalography and clinical neurophysiology, 43*(3), 362-370.