Studies in Second Language Acquisition (2022), 1–25
doi:10.1017/S027226312200047X

METHODS FORUM

Domain-general auditory processing as a conceptual and measurement framework for second language speech learning aptitude: A test-retest reliability study

Kazuya Saito[1]* and Adam Tierney[2]

[1]University College London, London, UK; [2]Birkbeck, University of London, London, UK
*Corresponding author. E-mail: k.saito@ucl.ac.uk

(Received 14 March 2022; Revised 19 October 2022; Accepted 27 October 2022)

Abstract

This article proposes a conceptual and measurement framework for postpubertal, L2 speech learning aptitude that is centered around domain-general auditory processing (i.e., representing spectral and temporal characteristics of sounds). To this end, we examine the construct and reliability of a battery of auditory processing tests by presenting the results of an empirical study wherein 100 participants took the tests twice. The findings showed (a) that the tasks tapped into multiple dimensions of auditory processing ability (e.g., perceptual acuity, audio-motor integration); and (b) that test-retest reliability ranged from "fair" to "excellent" (ICC $_{(2,2)}$ = .4–.8). Last, we introduce an offline version of the tests (deposited at *L2 Speech Tools for Researchers & Teachers*; http://sla-speech-tools.com/), provide a brief user manual, and invite researchers to use these tests to examine the role of auditory processing in various aspects of L2 speech learning (naturalistic vs. classroom, perception- vs. production-based practice).

Introduction

Auditory processing is the *perceptual* ability to precisely encode and integrate the spectral and temporal characteristics of sounds. Because the primary source of language input for most learners is the auditory channel, it has been proposed to serve as a bottleneck for first language (L1) acquisition (Goswami, 2015). Recently, scholars have begun to argue that auditory processing can uniquely explain individual variance in various aspects of second language (L2) speech learning (Mueller et al., 2012). Given that the auditory precision account of L2 speech learning is an emerging topic, it is important to call for more follow-up and replication studies with rigorous research designs. In this article, we first provide a *brief* introduction of the kinds of tasks used in cognitive psychology literature involving the two-way model of auditory processing—that is, types of processing skills (acuity, integration) × information content (spectral,

temporal). To help second language acquisition (SLA) researchers incorporate the auditory processing tasks and maximize the value of their research programs, we will then highlight one crucial issue related to methodological rigor—that is, the test-retest reliability of the auditory processing tests. To this end, we present the results of an empirical study wherein a total of 100 L1 and L2 speakers of English with varied biographical backgrounds took the same tasks twice. Finally, we introduce an offline version of the test batteries that researchers, students, and practitioners can use to assess auditory processing themselves and offer a range of directions for future auditory processing research.[1]

What is auditory processing and how is it measured in L1 acquisition literature?

Individuals differ greatly in the precision of their auditory perception (Kidd et al., 2007). Over the past 50 years, the relationship between auditory processing and language learning has been extensively researched in the field of cognitive psychology. Because auditory processing is considered one of the abilities that we first rely on when we encounter and parse aural input, and it is related to every dimension of linguistic processing (phonology, vocabulary, morphology), individual differences in this ability may impact the rate and ultimate attainment of language learning. Ample cross-sectional and longitudinal research demonstrates that this ability is linked to L1 development and impairment, although the causal nature of this link continues to be debated (for comprehensive overviews on the auditory precision account of L1 acquisition, see Goswami, 2015; Tallal, 2004; Tierney & Kraus, 2014; Wright et al., 1997; but for counterarguments, see Ramus, 2003; Rosen, 2003).

We define auditory processing here as the bottom-up, domain-general perceptual ability to discriminate and reproduce patterns along individual acoustic dimensions, such as pitch, formants, duration, and amplitude. Auditory perception is distinct from *speech* perception, in which sound categories are conveyed using multiple redundant sound characteristics (e.g., 18 different acoustic cues conveying the distinction between voiced and unvoiced consonants; Lisker, 1986). As with speech perception, a range of other phenomena (e.g., music, emotion, and environmental sounds) can also be perceived and categorized by accessing and weighting acoustic signals in different ways. This corresponds to an influential view in cognitive psychology that states that auditory processing can be considered domain-general and this ability forms the basis of multiple domain-specific phenomena (for a comprehensive overview, see Mueller et al., 2012).

[1]Importantly, the primary objective of this article is to introduce SLA researchers to the existing paradigms in cognitive psychology (the auditory precision account of language learning) and relevant task formats (i.e., discrimination for acuity, reproduction for integration) rather than propose entirely new theoretical and methodological frameworks of our own. In addition, the focus of the article concerns the extent to which the tasks can elicit similar enough scores when participants do them twice (i.e., test-retest reliability); however, we do *not* explore the extent to which the tasks tap into what they aim to measure (i.e., construct validity) as this topic has still been discussed in the field of cognitive psychology. To date, much attention has been directed toward examining precisely what characterizes auditory processing relative to other cognitive abilities (e.g., Snowling et al., 2018) and how such abilities can be measured by using both behavioral and neural instruments and comparing their results (e.g., Clinard et al., 2010). For a more detailed overview on the theoretical underpinnings of auditory processing and the construct validity of the tasks in the context of L2 speech learning, see Saito, Suzukida et al. (2021).

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

To measure auditory processing abilities domain-generally, tasks typically employ a number of synthesized, nonverbal stimuli with very simple acoustic characteristics (e.g., completely flat fundamental frequencies, formant contours, and harmonic spectrum) that normal hearing listeners will not perceive as speech. The stimuli are all identical except for one acoustic dimension (e.g., pitch, duration). During the tasks, participants are asked to discriminate between or reproduce the stimuli based on the target acoustic dimension.

Auditory processing can be subdivided in several different ways. In this article, we introduce the two-way auditory processing model to help SLA researchers understand the conceptual and methodological paradigms in cognitive psychology and L1 acquisition research and incorporate them into their L2 acquisition research. A first key distinction concerns the type of processing skills—the encoding of subtle acoustic characteristics of sounds (perceptual acuity; Moore, 2012), and the use of acoustic information for motor action (audio-motor integration; Corriveau & Goswami, 2009; Tierney et al., 2017). Another key distinction is the type of acoustic information that speakers process, for example, spectral information (pitch and formant frequency) or temporal information (duration and amplitude rise time) (Kidd et al., 2007).

The perceptual acuity component of auditory processing is defined as one's capacity to precisely encode the acoustic details of sounds. It is often measured using psychoacoustic discrimination tasks, in which learners are asked to discriminate between two or three complex, nonspeech sounds. In this task, all acoustic properties of the stimuli except for one are identical so that researchers can measure participants' sensitivity to a single cue (e.g., pitch, formant, duration, or amplitude). A standard stimulus is followed by one or two comparison stimuli (AX or AXB discrimination format). The objective of the task is to examine how small of a difference participants can detect between the standard versus comparison stimuli. The presentation of the stimuli can occur in set blocks (e.g., Test of Basic Auditory Capabilities; Kidd et al., 2007) or through an adaptive procedure in which the level of difficulty changes per trial depending on participants' performance (e.g., Leek, 2001).

Precise auditory processing (i.e., fine-grained analyses of aural input) enables listeners to make fine perceptual distinctions between phenomena (perceptual acuity), but auditory processing can comprise a range of neighboring abilities beyond perceptual acuity (e.g., Tierney & Kraus, 2014 for the precise auditory timing hypothesis). In fact, simple discrimination of sounds does not do justice to the complex ways in which language learners interact with sound while learning to produce and perceive speech (Liberman & Mattingly, 1985). For example, neuroimaging studies have revealed that speech perception draws upon motor planning resources, even when participants are instructed not to move (Pulvermuller et al., 2006). This interaction between auditory and motor resources is not limited to linguistic stimuli; perception of nonverbal rhythms, for example, consistently lead to activation in cortical and subcortical motor areas such as the basal ganglia and supplementary motor area (Grahn & Brett, 2007). Auditory processing is not, therefore, a purely perceptual phenomenon, but requires integration of perceptual and motor resources, regardless of whether a listener's task is to categorize a stimulus or to repeat it back.

Therefore, auditory processing can be reconceptualized as a set of perceptual *and* cognitive abilities. Another component of auditory processing, therefore, is audio-motor integration, defined as the capacity to capture basic acoustic characteristics of sounds and use them to modulate motor action. Individuals can show different levels of acuity and integration. While some can perceive very subtle acoustic differences, they

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

may have difficulty proceduralizing them (high acuity, low integration). Accordingly, acuity and integration were found to follow different developmental trajectories, suggesting that they reflect two essentially different aspects of auditory processing: While acuity tends to reach its peak around puberty and then continuously degrades, integration continues to improve until the late twenties (Thompson et al., 2015).

One way to operationalize audio-motor integration is through a reproduction task, in which learners are asked to listen to and replicate a set of target sound sequences (Flaugnacco et al., 2014). For spectral integration, participants use a piano-like keyboard with five notes to play back melodic sequences (e.g., Saito, Suzukida et al., 2021 for a melody reproduction task). For temporal integration, participants use a drum to repeat back rhythm sequences (Tierney et al., 2014 for a rhythm production task; see also Dellatolas et al., 2009). Other scholars have measured audio-motor integration using a synchronization paradigm, wherein participants are asked tap in time with rhythmic sequences generated by a metronome (e.g., Corriveau & Goswami, 2009; Surányi et al., 2009).

How is auditory processing relevant to L2 speech learning?

In the context of adult L2 speech learning, there is a growing amount of evidence showing that different types of auditory processing (acuity, integration) are uniquely associated with different aspects of L2 speech learning (accuracy vs. fluency, segmentals vs. suprasegmentals, phonology vs. lexicogrammar; for comprehensive reviews, see Saito, in press; Saito et al., 2021). Studies have shown that those with more precise perceptual acuity are able to better perceive and learn foreign sounds to which they have not previously been exposed (e.g., Kempe et al., 2015; Chandrasekaran et al., 2010; Perrachione et al., 2011; Wong & Perrachione, 2007). The predictive power of acuity is most clearly observed with naturalistic L2 learners who have acquired a target language through communicatively authentic, immersive experiences. Specifically, it is a medium-to-strong predictor of L2 speech learning outcomes for those with medium-to-long-term immersion experience (length of residence = 1–10 years; Kachlicka et al., 2019; Saito et al., 2020), but not for those with short-term immersion experience (length of residence < 6 months; Saito, Sun et al., 2022) or for foreign language learners (length of residence = 0 years; Saito et al., 2021; for longitudinal evidence on the role of acuity in the first three months and one year of immersion, see Sun et al., 2021; and Saito et al., 2020a). Additionally, training studies have shown that those with greater acuity demonstrate more improvement when they are intensively exposed to target sounds produced by multiple speakers under various phonetic and lexical conditions (i.e., high variability phonetic training; Lengeris & Hazan, 2010 for English vowels; Qin et al., 2021 for Cantonese lexical tones).

Audio-motor integration has been linked to different aspects of L2 speech proficiency, especially when learners prioritize the production, repetition, and proceduralization of language. For instance, in naturalistic settings, integration has been associated with both the fluency and prosodic accuracy aspects of L2 speech learning (Saito, Kachlicka et al., 2020 for fluency; Sun et al., 2021 for prosodic accuracy). Other evidence suggests that integration can also be predictive of speech learning outcomes in foreign language classrooms. In such contexts, L2 learners' practice opportunities are typically restricted to several hours of language-focused instruction per week, wherein most activities are production-based (e.g., Shintani et al., 2013). When it comes to speaking, one common production-based activity is choral repetition of teachers'

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

model pronunciation forms (e.g., Baker, 2014). Few foreign language learners are able to access ample communicatively authentic input—a key element for the development, refinement, and sophistication of their L2 representational systems. In this context, integration (rather than acuity) appears to be a driving force of successful learning (Saito et al., 2021). Longitudinal evidence has also shown that integration is a key factor of L2 acquisition when learners are engaged in production-based speech training (e.g., Li & DeKesyer, 2017 for word reading and picture description; Shao et al., 2022 for shadowing).

Finally, different types of information processing (spectral vs. temporal) can also be uniquely associated with different aspects of L2 speech learning. According to Saito and Plonsky's (2019) model of L2 speech proficiency, it comprises segmental abilities (i.e., the accurate perception and production of individual sounds), melodic and prosodic abilities (i.e., the varied and appropriate use of word and sentence stress, and intonation), and temporal and fluency abilities (i.e., optimal delivery without too many pauses and repetitions). On the one hand, spectral processing has been found to predict the development of segmental abilities (e.g., Lengeris & Hazan, 2010 for the link between formant acuity and vowel acquisition) and melodic and prosodic abilities (e.g., Qin et al., 2021 for the link between pitch acuity and lexical tone acquisition). On the other hand, temporal processing has been linked to L2 fluency acquisition (e.g., Saito et al., 2020a for the link between duration processing and fluency development and attainment).

Current study

So far, we have provided a brief overview on the model of auditory processing (acuity, integration) and measurement practices (discrimination, reproduction) in the field of cognitive psychology and SLA. However, few studies have explored the extent to which the relevant tasks (discrimination, reproduction) can *reliably* tap into the acuity and integration constructs of auditory processing. To test the reliability of the representative measurement formats, here we present the results of a study wherein a total of 100 adult L1 and L2 English speakers took each task twice (three discrimination tasks for acuity, two reproduction tasks for integration). As described in more depth in the following text, for the discrimination tasks we followed Levitt's (1971) procedure by making the tests *individually adaptive*: Every participant engaged in different stimuli specially selected to ensure that the task was as difficult as could be without being impossible. This was done to minimize the amount of testing time and increase the precision of the analyses. Because the tests were internally adaptive, this article highlighted the analyses of test-retest reliability; the internal consistency of the test stimuli was not pursued because the design of the tests did not allow us to explore participants' performance on the same stimuli. We aimed to answer the following questions:

1. What are the latent factors underlying auditory processing as measured using discrimination and reproduction tasks?
2. What is the test-retest reliability of the auditory processing tasks?

At the end of this article, we will introduce and discuss an open-source, freely available auditory-processing test battery to help L2 researchers measure auditory processing in their own research programs.

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Method

Participants

A total of 100 participants were recruited to engage in the tasks under two different conditions ($n = 54$ for lab; $n = 46$ for online). The biographical backgrounds of the participants are summarized in Table 1. Participants in the laboratory group ($n = 54$) were college-level students enrolled in English-as-a-Foreign-Language classes at a university in Spain (5 males, 49 females). The participants were recruited as a part of a larger project that examined the relationship between Spanish EFL learners and the outcomes of L2 English learning. On Day 1 (T1), participants completed the tasks in a quiet room in tandem with a researcher at the university. Four days later, they returned to the lab and took the same tests again (T2).

In addition to the laboratory group, 46 participants were recruited. We initially planned to collect the data in face-to-face settings. Due to the ongoing COVID-19 situation and practical constraints, however, this part of the data collection was conducted using the online psychology experiment platform GORILLA (Anwyl-Irvine et al., 2020). Participants in the online group included native speakers of Chinese ($n = 24$), Japanese ($n = 8$), English ($n = 5$), Polish ($n = 4$), French ($n = 2$), and other Indo-European languages ($n = 3$). It is important to acknowledge that the two groups (lab vs. online) were not comparable in many different ways (e.g., L1 backgrounds, L1/L2 use; see Table 1); and that the comparison of the lab and online groups was not a part of our original research questions.

To recruit participants with varied biographical backgrounds, an electronic flyer was created, posted, and disseminated on a range of social media websites (e.g., Twitter, Facebook). We clarified that we were looking for participants without any previous hearing problems. Interested participants contacted the investigation team using e-mail. None of them reported any hearing impairment.

Each participant was introduced to the main objective of the study and the procedure for the auditory processing tests. Given that participants resided in different time zones, they were allowed to complete the tests according to their own schedule but following a suggested schedule: T1 and T2 with a flexible interval of 1–4 days. To examine the test-retest performance, participants completed the auditory processing tests using the same procedure (described in the text that follows). The progress and timing of the participants' performance was tracked using the Gorilla platform. Participants were also encouraged to contact a member of the investigation team if they had questions regarding the auditory processing tasks and/or encountered problems with the test-taking procedures.

Prior to the project, we had accumulated ample experience in conducting auditory processing tests using the Gorilla platform. To elicit high-quality data, participants were asked to use their laptop computer (rather than a smartphone or a tablet) with

Table 1. Demographical backgrounds of 100 participants

	A. Laboratory Group ($n = 54$)			B. Online Group ($n = 46$)		
	M	*SD*	*Range*	*M*	*SD*	*Range*
Age	20.5	4.6	19–53	27.4	8.1	18–49
L2 English use[a]	24.4%	15.1	5–70%	43.0	36.8	1–100%
EFL vs. ESL[a]	EFL = 54, ESL = 0			EFL = 24, ESL = 17		

[a]Excludes five native speakers of English from the online group.

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

headphones, use Google Chrome, and complete all the tasks in a quiet room with a stable internet access.

In the current manuscript, the main objective of the project concerns the degree of reliability of the auditory processing test battery when participants took them twice (i.e., test-retest reliability). Participants in this study included both L1 and L2 English speakers with diverse biographical backgrounds. To our knowledge, there has been no discussion, evidence, or controversy surrounding whether the strength of the test-retest reliability may differ among certain groups of participants (e.g., advanced L2 speakers vs. beginner L2 speakers vs. L1 native speakers; young vs. old participants; tonal vs. non-tonal language users; musicians vs. non-musicians). Thus, we treated all the participants as one single group, and did not further conduct any analyses on the test-retest reliability as per participants' biographical and L1/L2 proficiency backgrounds.

Auditory processing test batteries

In accordance with the two-way auditory processing model, a total of five tasks were created:

- Formant discrimination (acuity/spectral);
- Pitch discrimination (acuity/spectral);
- Duration discrimination (acuity/temporal);
- Melody reproduction (integration/spectral); and
- Rhythm reproduction (integration/temporal).

The discrimination tasks were designed to assess participants' sensitivity to one particular acoustic dimension of nonverbal sounds (formants, pitch, and duration). The tasks follow key methodological features in other discrimination tasks in the cognitive psychology literature, that is, AXB discrimination of nonspeech stimuli (Moore, 2012). Although some tests ask participants to work on the same set of stimuli (e.g., Test of Basic Auditory Capabilities; Kidd et al., 2007), our test adopted Levitt's (1971) individually adaptive procedure to enhance the precision of the tests and reduce the amount of testing time (e.g., Leek, 2001).

The reproduction tasks were designed to assess participants' capacity to capture patterns over time within an acoustic dimension (pitch or duration) and integrate them into motor action. For the sake of comparability, our reproduction tests followed the methods in the existing literature, that is, Tierney et al. (2017) for rhythm reproduction and Saito et al. (2021) for melody reproduction.

The five tasks in the current study have already been used in previous literature showing that the resulting scores could predict various dimensions of L2 speech learning (e.g., immersion vs. classrooms, early vs. late phase of immersion, phonology vs. lexicogrammar aspects of speech). The tasks were developed in-house, but the formats are essentially comparable to those of discrimination and reproduction tasks widely used in cognitive psychology.

During each session (Days 1 and 2), participants completed the auditory processing tests in the following order: (1) formant discrimination (3–5 min), (2) pitch discrimination (3–5 min), (3) duration discrimination (3–5 min), (4) melody reproduction (5 min), and (5) rhythm reproduction (5 min). Given that the tests were designed to measure auditory precision, the participants in the online group were explicitly told to

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

do the tasks in a quiet room and that doing the tests under noisy conditions would negatively affect the reliability of the results. Those in the lab group completed the tasks in a soundproof room.

Discrimination (formant, pitch, and duration)

Stimuli

Following the psychoacoustic literature (e.g., Test of Basic Auditory Capabilities; Surprenant & Watson, 2001. For a methodological summary, see Moore, 2012), the discrimination task comprised three subtests for the following acoustic dimensions: formant, pitch, or duration. In each subtest, 101 nonverbal sounds were prepared (1 standard stimulus and 100 comparison stimuli). These sounds were synthesized such that they differed only in formant frequency (F2 = 1500–1700 Hz with a step of 2 Hz), pitch (F0 = 330–360 Hz with steps of 0.3 Hz), or duration (250–500 ms with steps of 2.5 ms). The standard stimulus was labeled as Level 0, and the comparison stimuli as Levels 1–100. Each level corresponded to distance in a target acoustic continuum (e.g., Level 1 for 1502 Hz and Level 2 for 1504 Hz in formant discrimination). By using the AXB discrimination format, we were able to examine the smallest difference between the standard and comparison stimuli that participants could hear. The standard stimulus was always the same while the comparison stimulus could change at each trial as per participants' performance (e.g., Level 0 vs. Level 1, Level 0 vs. Level 30, Level 0 vs. Level 60).

- **Formant discrimination:** Each sample had a duration of 500 ms with two 15 ms amplitude ramps at the beginning and endpoint of the stimulus. The fundamental frequency was set to 100 Hz with harmonics up to 3000 Hz. Three formants were inserted at 500 Hz, 1500 Hz, and 2500 Hz. The F2 of the standard stimulus was set to 1500 Hz (Level 0) using a parallel formant filter bank (Smith, 2007). The F2 of the comparison stimuli ranged from 1502 to 1700 Hz at a step size of 2 Hz (Levels 1–100).[2]
- **Pitch discrimination:** Four-harmonic complex tones were prepared, with the fundamental frequency set to 330 Hz and with equal amplitude across harmonics. The duration of each stimulus was 250 ms. F0 was set to 330 Hz for the standard stimulus (Level 0) but ranged from 330.3 Hz to 360 Hz for the comparison stimuli with increments of 0.3 Hz (Levels 1–100).
- **Duration discrimination:** The same four-harmonic tones from the pitch discrimination task were used for the duration discrimination task. Although F0 remained constant at 330 Hz, the duration of the standard stimulus was set to 250 ms (Level 0) and the duration of the comparison stimuli changed from 252.5 ms to 500 ms with increments of 2.5 ms (Levels 1–100). For a summary of the standard versus comparison stimuli, see Table 2. All the sound samples used for the discrimination tasks can be found in Supporting Information-A and in IRIS.

[2]In the prior study (Saito, Kachlicka, Suzukida, Petrova et al., 2022), 10 human listeners rated the speech-likeness of the nonspeech stimuli (together with speech stimuli) on a 10-point scale (10 = Definitely human speech, 1 = not human speech-like at all). The nonverbal stimuli were perceived as "not human speech-like at all" (M = 2.02–2.12 out of 10).

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Table 2. Summary of stimuli in discrimination tasks

Measures	Standard stimulus (Level 0)	Comparison stimuli (Levels 1–100)	Step size
Formant discrimination	1500 Hz	1502–1700 Hz	2 Hz
Pitch discrimination	300 Hz	330.3–360 Hz	0.3 Hz
Duration discrimination	250 ms	252.5–500 ms	2.5 ms

General procedure

For each subtest, three stimuli were presented per trial. The second sound was always the standard stimulus. Following the format of the AXB discrimination task, either the first or the third sound was different from the other two.[3] If the order was Comparison-Standard-Standard, the first one was different. If the order was Standard-Standard-Comparison, the last one was different. Participants were asked to indicate which sound was different by either pressing the number "1" or "3" on a keyboard. For task instruction and onscreen labels that participants accessed during the tasks, see Figure 1.

A total of 101 stimuli (i.e., one standard stimulus [Level 0] and 100 comparison stimuli [Levels 1–100]) were presented to participants. If a participant could detect the difference between Level 0 and Level 1, their perceptual acuity could be considered extremely precise for the target acoustic domain. If the minimum difference was Level 100, however, their perceptual acuity could be considered low (indicating that Levels 0–99 may sound very much the same to them). In this regard, the finer the levels participants can discriminate, the more precise their perceptual acuity.

The tests started from the mid-point, Level 50. Based on Levitt's (1971) adaptive procedure, the difficulty of the task changed with each trial. If a participant selected an incorrect response, the difficulty of the task decreased by 10 levels (e.g., Levels 50 → 60, the difference between comparison stimuli becoming wider). When a participant made three consecutive correct responses, the task difficulty increased by a degree of 10 levels (e.g., Levels 50 → 60, the difference between comparison stimuli reducing).

The step size decreased when the direction of a participant's performance reversed, that is, incorrect answers after a string of correct answers, or correct answers after a string of incorrect answers. The step size changed from 10 to 5 at the first reversal, and

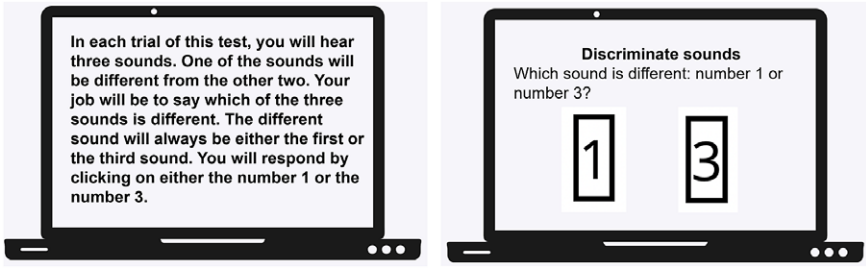

Figure 1. Task instruction and onscreen labels: Discrimination task.

[3]While AXB commonly refers to a task where participants determine whether A or B matches X, we use this term (AXB) in a different sense. In our task, participants are asked to determine whether A or B is *different*, meaning that this is an inverse version of AXB discrimination (e.g., if A is different, B matches X).

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

from 5 to 1 after the second reversal. As such, the task was able to ultimately measure a participant's discrimination threshold very precisely.

The program continued after either 70 trials were completed, or until eight reversals were reached. The threshold was then calculated using average of the levels of each reversal from the third onward. For example, if a participant reached eight reversals at Levels 60, 55, 56, 45, 47, 38, 45, and 42, their score would be calculated as the average of the last six numbers (i.e., 47.1 out of 100). This would suggest that a participant could barely hear the difference between Levels 0 and 47. Stimulus level was set at different rates for each subtest (1 level = 2 Hz for formant, 0.3 Hz for pitch, and 2.5 ms for duration).

For the subsequent analyses, we used subcomponent scores (formant, pitch, and duration) and overall scores (standardized and averaged scores).

Reproduction (melody, rhythm)

Sixteen participants from the online group did not take the reproduction task for logistical reasons. In addition, six lab participants' rhythm reproduction scores were not properly recorded due to technological issues. Thus, in the current study, the data of 84 participants were analyzed for melody reproduction ($n = 54$ for lab; $n = 30$ for online), and that of 78 participants for rhythm reproduction ($n = 48$ for lab; $n = 30$ for online). Building on Tierney et al. (2017), participants were asked to listen, remember, and repeat sequences of sounds that varied (a) in fundamental frequency for melody reproduction; and (b) in the ratio of drum hits and rests for rhythm reproduction. Tracking an entire sound sequence across time requires spectral and temporal processing on a slow time scale (i.e., 2 seconds). Tracking sound sequences on such slow time scales may require integration between auditory regions and motor planning regions in the brain (Patel & Iverson, 2014; Tierney et al., 2017).

There were a total of 10 melodies (300 ms per note) in the melody reproduction task. Each melody consisted of a sequence of seven notes. Each note comprised a set of five six-harmonic complex tones, with amplitude held constant across harmonics. Their fundamental frequencies were set to 220, 246.9, 277.2, 311.1, and 329.6 Hz, which correspond to the first five notes of the A major scale. Each melody always began on the third note of the scale (i.e., 277.2 Hz). The next note was randomly chosen to be either one note higher on the scale (311.1 Hz) or one note lower on the scale (246.9 Hz), so there is never an interval of more than one note. This process then repeated up to the seventh note. Once the melody reached the lower (220 Hz) or upper end (329.6 Hz), the next note was chosen to either be closer to the center of the range (246.9 Hz or 311.1 Hz) or identical to the previous note (220 Hz or 329.6 Hz).

At the beginning of the test, participants were shown a set of five buttons that were arranged in a line stretching from the top to the bottom of the screen. They were encouraged to try clicking the buttons to hear the tone that each button produced. Higher frequency tones were linked to buttons that were closer to the top of the screen. To examine how they could adjust to the *new* motor task (i.e., hitting the buttons in response to sound prompts), and to minimize the influence of their prior relevant experience (e.g., piano training), the buttons were displayed vertically rather than horizontally (the latter simulates piano playing).

During each trial, participants listened to the same sequence repeated three times, and then repeated it by pressing the five buttons. To reduce participants' memory load, each sequence was repeated three times (cf. sound and letter sequences are displayed

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

only once in working memory tests). To examine how quickly participants could integrate the received sound input into action, they were not allowed to relisten to a sequence more than three times.

The accuracy ratio (%) was recorded based on the first seven button presses. Notes were scored as a 1 when they were identical to target notes and scored as a 0 when they differed from target notes to any degree. Performance was then averaged across all 10 melodies. For task instruction and onscreen labels that participants accessed during the tasks, see Figure 2.

The rhythm reproduction task used a total of 30 rhythmic patterns generated from Povel and Essens's (1985) notion of strongly versus weakly metrical sequences (3.2 seconds per token). The strongly metrical sequences contain more drum hits on the first and third beats than weak metrical sequences. The drum hits consisted of a 150-ms conga drum hit sound acquired from www.freesound.org. After listening to the stimuli, the participants were asked to reproduce the rhythm by repeatedly pressing the space key. Unlike the melody reproduction task (which was novel to all participants), we assumed that the format of the rhythm reproduction task (i.e., drumming beats) was familiar thus minimizing the influence of practice and experience effects (but see the relationship between rhythm integration and music experience; Tierney & Kraus, 2014).

The interpress times were first quantized by changing them to the nearest interval in the set [200 400 600 800 1000 1200 1400 1600 1800 2000 ms]. Participants' responses were then traced as a sequence of hits and rests, which were then compared to the sequence of hits and rests in each stimulus. Accuracy ratio (%) was recorded in terms of the presence of hits or rests at every 200 ms interval. For task instruction and onscreen labels that participants accessed during the tasks, see Figure 3. All melodic and rhythmic sequences can be found in Supporting Information-A and in IRIS. For the

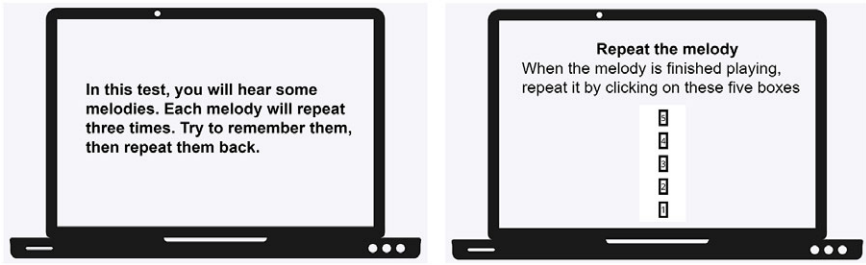

Figure 2. Task instruction and onscreen labels: Melody reproduction task.

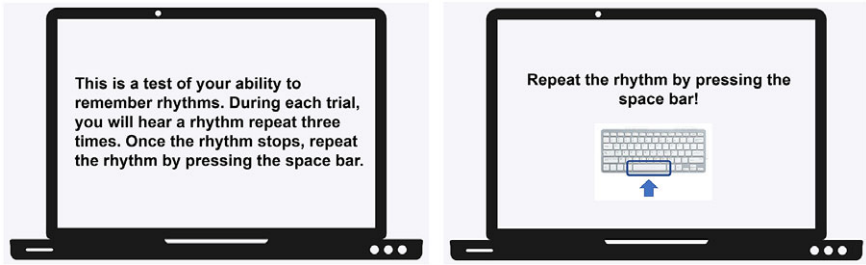

Figure 3. Task instruction and onscreen labels: Rhythm reproduction task.

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

subsequent analyses, we used subcomponent scores (melody, rhythm) and overall scores (standardized and averaged scores).

Results

A set of factor analyses were performed to examine to what extent the participants' performance elicited through five tasks reflected the two-way model of auditory processing (spectral/acuity, spectral/integration, temporal/acuity, temporal integration).

Components of auditory processing

For the reasons mentioned in the preceding text, whereas all participants took the discrimination tests, 78 completed the reproduction tests. Two separate factor analyses were performed to examine the presence of latent variables underlying their discrimination scores (n = 100) and reproduction scores (n = 78). Finally, we performed another factor analysis based on the 78 participants who completed both acuity and reproduction tasks.

Components of discrimination scores

As summarized in Supporting Information-B, n = 100 participants' perceptual acuity scores (the smallest difference they can discriminate in Hz and ms) widely varied at T1 (first test) and T2 (second test). The results of Kolmogorov–Smirnov tests showed that the scores did not significantly differ from normal distributions at T1 and T2 ($p > .05$). In terms of the constructs of the discrimination test, their subcomponent discrimination scores (pitch, formant, and duration) at T1 and T2 were submitted to an exploratory factor analysis with Varimax rotation to identify any patterns underlying the multiple discrimination test measures.

To ensure the robustness of the factor analyses, the appropriateness of the sample size was examined carefully. Following the guidelines set by Loewen and Gonulal (2015), the size of the current study could be considered adequate for the use of factor analyses. First, the total number of participants (N = 100) met Hair and colleagues' (1998) suggestions for minimum sample sizes. Second, as reviewed in the preceding text, the constructs of perceptual acuity (measured by the discrimination tasks) can be assumed to comprise two distinct abilities—spectral (formant, pitch) and temporal/acuity (duration). Thus, the number of participants per predicted variable (i.e., n = 50 for two different aspects of auditory processing respectively) was beyond Field's (2009) threshold (n = 10–15). Finally, the KMO measure of sampling adequacy yielded .725; the value here could be considered "good" as per Field's (2009) benchmarks (x > .5 for "mediocre"; x > .7 for "good"; x > .8 for "perfect").

In L2 research, given that the cumulative percentage of explained variance is typically small (e.g., 60%), Loewen and Gonulal (2015) recommended that the number of factors should be determined when the model explains at least 70% of the variances as a cut-off point. Thus, two latent factors were identified as shown in Table 3 (explaining 75.3%). The factorability of the entire dataset was confirmed through two tests: Bartlett's test of sphericity (χ^2 = 124.897, $p < .001$) and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (.725). To interpret the profiles of Factors 1, 2, and 3, 0.4 was used as a cut-off value for the "practically" significant factor loadings (Hair et al., 1998). The findings concurred with the assumptions in the two-way model

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Table 3. Summary of a two-factor solution based on a factor analysis of perceptual acuity score

	Factor 1: Spectral acuity	Factor 2: Temporal acuity
Cumulative%	40.8%	34.5%
Formant discrimination (T1)	**.734**	.249
Formant discrimination (T2)	**.634**	.281
Pitch discrimination (T1)	**.739**	−.014
Pitch discrimination (T2)	**.790**	−.023
Duration discrimination (T1)	−.046	**.851**
Duration discrimination (T2)	.276	**.774**

Note: All loadings > .4 were highlighted in bold.

of auditory processing that the three discrimination tasks related to the two distinct components of acuity, that is, formant and pitch discrimination for spectral acuity (Factor 1) versus duration discrimination for temporal acuity (Factor 2).

Components of reproduction test

As summarized in Supporting Information-B, Participants' audio-motor integration scores (how accurately they can reproduce melodic and rhythmic sequences) did not deviate significantly from normal distribution (Kolmogorov–Smirnov; $p > .05$). The participants' melody and rhythm reproduction task performance at T1 and T2 were submitted to a factor analysis with Varimax rotation. The factorability of the entire dataset was confirmed using two tests: the Bartlett's test of sphericity ($\chi^2 = 173.194$, $p < .001$) and the KMO measure of sampling adequacy (.610). A decision was made to identify a "two-factor" solution which accounted for 90.5% of the total variance in the participants' reproduction scores (see Table 4). As predicted earlier, their melody and rhythm reproduction scores at T1 and T2 were clustered into two different groups, respectively. Therefore, Factor 1 was labeled as spectral integration and Factor 2 as temporal integration.

Components of acuity and reproduction tests

Focusing on 78 participants who completed both discrimination and reproduction tasks, another factor analysis was performed with Varimax rotation. The factorability of the model was adequate: the Bartlett's test of sphericity ($\chi^2 = 337.434$, $p < .001$) and the KMO measure of sampling adequacy (.729). A "four-factor" solution was selected because the model accounted for 74.9% of the total variance in the entire dataset (see Table 5). As predicted, the three tasks were clustered to three different factors, suggesting that they may relate to three distinguishable abilities (Factor 2 for rhythm integration, Factor 3 for spectral acuity, and Factor 4 for temporal acuity). Unlike our

Table 4. Summary of a two-factor solution based on a factor analysis of audio-motor integration scores

	Factor 1: Spectral integration	Factor 2: Temporal integration
Cumulative%	62.7%	27.8%
Melody reproduction (T1)	**.912**	.280
Melody reproduction (T2)	**.955**	.108
Rhythm reproduction (T1)	.226	**.916**
Rhythm reproduction (T2)	.142	**.935**

Note: All loadings > .4 were highlighted in bold.

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Table 5. Summary of a four-factor solution based on a factor analysis of entire dataset

	Factor 1: Pitch processing	Factor 2: Rhythm integration	Factor 3: Spectral acuity	Factor 4: Temporal acuity
Cumulative%	41.3%	13.7%	10.6%	9.3%
Formant discrimination (T1)	−.341	−.252	**.648**	.025
Formant discrimination (T2)	−.178	−.169	**.876**	−.009
Pitch discrimination (T1)	**−.619**	−.364	−.011	.259
Pitch discrimination (T2)	**−.715**	−.170	.126	.020
Duration discrimination (T1)	.018	−.112	−.007	**.904**
Duration discrimination (T2)	−.108	−.012	**.600**	**.573**
Melody reproduction (T1)	**.825**	.197	−.236	.081
Melody reproduction (T2)	**.853**	.033	−.200	.014
Rhythm reproduction (T1)	.184	**.872**	−.196	−.016
Rhythm reproduction (T2)	.222	**.897**	−.162	−.143

prediction, however, the results showed that both pitch discrimination and melody reproduction were reduced into the same factor. This in turn indicated that the two tasks may tap into the same aspect of pitch processing (Factor 1).

Test-retest reliability

A set of test-retest reliability analyses were performed to explore whether the discrimination and reproduction tasks could assess a stable construct of auditory processing abilities. Traditionally, the Pearson's product-moment correlations between the test and the retest scores were evaluated. However, this approach could be problematic because participants' scores are not independent when they take the same tests twice. Following Liljequist et al.'s (2019) recent model of test-retest reliability, we thus decided to use intraclass correlations (ICCs) that take into account the repeated nature of the data and the existence of systematic errors (bias). In the case of the current study, there can be multiple sources of error, such as within-participant variation (i.e., individuals demonstrating different test performance) and test-retest variation (i.e., participants' differing performance between the first and second tests). Thus, a two-way random model was selected. Because participants took the auditory processing tests only once at each testing point, a single (rather than average) measure ICC was used.

ICC can be conceptualized as the ratio of variance of interest to total variance. If the variance of the interest is higher than the unwanted variance, the reliability of the instruments can be considered excellent. To control for the magnitude of subject errors (e.g., individual differences) and measurement errors (e.g., two testing points), and to reflect the relationship between the same participants' test performance at two different time points (T1 vs. T2), a set of ICC analyses were performed. As Liljequist et al. (2019) suggested, two different types of ICC values were provided—that is, consistency ICCs (with biases) versus absolute agreement ICCs (without biases)—to allow readers to see the reliability of the test materials with and without biases. If there were biases, consistency ICCs could be higher than absolute agreement ICCs. To check the significance of the differences between consistency and absolute ICCs, 95% confidence intervals were used. Whereas a range of guidelines have been suggested (e.g., Liljequist et al., 2019), it is important to interpret ICC values as per research context. Thus, we decided to use Cicchetti's (1994) field-specific benchmark for assessment tools in the field of psychology: 0.4 to 0.6 for "fair," 0.6 to 0.75 for "good," and 0.75+ for "excellent."

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Table 6 summarized both consistency and absolute agreement ICC coefficient values. According to 95% CI analyses, participants' consistency and absolute agreement ICCs overlapped to a great degree, suggesting that the influence of biases may be minimal in the current dataset (Liljequist et al., 2019). For the rest of the analyses, therefore, we focused on absolute agreement ICCs (without biases). As for participants' subcomponent scores, although the ICC values were lower for duration discrimination compared to formant and pitch discrimination [ICC $_{(2,2)}$ = .409 vs. .520/.525], they could all be considered as "fair." The ICC values of the reproduction subtests were good to excellent [ICC $_{(2,2)}$ = .796 for melody reproduction and .744 for rhythm reproduction].

In some previous studies (e.g., Saito, Sun et al., 2022), scholars have operationalized participants' *overall* auditory processing of various dimensions of acoustic signals. Thus, we calculated overall discrimination scores by standardizing and averaging the formant, pitch, and duration scores, and overall reproduction scores by standardizing and averaging the melody and rhythm scores. The ICC values were considered good for overall discrimination [ICC $_{(2,2)}$ = .625 for absolute agreement] and excellent for reproduction [ICC $_{(2,2)}$ = .843 for absolute agreement].

Post-hoc analyses: lab versus online

Certain parts of the data (n = 46 out 100) were collected online due to the pandemic situation and practical constraints. Although the two subgroups of the participants differed in various aspects (e.g., L1 background, L2 proficiency, immersion experience), and comparing the role of testing conditions (lab vs. online) was not originally conceptualized in the research design, we conducted a set of post-hoc analyses to test whether the ICC values differed between the laboratory group (n = 54) and the online

Table 6. Summary of intraclass correlation coefficients

		Intraclass coefficients (Single measures)	95% CI			
			Lower	Upper	Value	*p*
Overall discrimination	Consistency	.622	.486	.729	4.295	< .001
	Absolute agreement	.625	.488	.731	4.295	< .001
Formant discrimination	Consistency	.571	.422	.689	3.659	< .001
	Absolute agreement	.520	.302	.674	3.659	< .001
Pitch discrimination	Consistency	.541	.387	.666	3.361	< .001
	Absolute agreement	.525	.363	.655	3.361	< .001
Duration discrimination	Consistency	.412	.234	.562	2.387	< .001
	Absolute agreement	.409	.233	.560	2.387	< .001
Overall reproduction	Consistency	.841	.762	.896	11.591	< .001
	Absolute agreement	.843	.764	.897	11.591	< .001
Melody reproduction	Consistency	.826	.744	.884	10.515	< .001
	Absolute agreement	.796	.638	.879	10.515	< .001
Rhythm reproduction	Consistency	.782	.678	.855	8.176	< .001
	Absolute agreement	.744	.555	.848	8.176	< .001

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Table 7. Summary of consistency intraclass correlation coefficients under lab versus online conditions

	Consistency intraclass coefficients (Single measures)	95% CI			
		Lower	Upper	Value	*p*
A. Laboratory Group ($n = 54$)					
Overall discrimination	.689	.517	.807	5.341	< .001
Formant discrimination	.470	.181	.672	3.302	< .001
Pitch discrimination	.563	.344	.723	3.840	< .001
Duration discrimination	.457	.217	.645	2.387	< .001
Overall reproduction	.839	.730	.906	11.642	< .001
Melody reproduction	.727	.559	.836	6.889	< .001
Rhythm reproduction	.795	.649	.882	8.176	< .001
B. Online Group ($n = 46$)					
Overall discrimination	.500	.246	.690	2.960	< .001
Formant discrimination	.533	.282	.714	3.599	< .001
Pitch discrimination	.436	.172	.643	2.549	< .001
Duration discrimination	.372	.096	.595	2.182	.005
Overall reproduction	.841	.683	.922	12.677	< .001
Melody reproduction	.814	.406	.928	15.113	< .001
Rhythm reproduction	.670	.245	.853	7.145	< .001

group ($n = 46$). The results of intraclass correlation analyses were summarized in Table 7. Due to the exploratory nature of the data, the interpretations here can be considered "tentative."

In terms of participants' overall performance, the ICC values of discrimination scores were fair in the online group [$ICC_{(2,2)} = .500$] but good in the laboratory group [$ICC_{(2,2)} = .689$]; the test-retest reliability of reproduction was comparable [$ICC_{(2,2)} = .839$ for laboratory and .841 for online]. The ICC values of the subcomponent scores were fair to excellent across the different group conditions (.4 to .8). One notable difference was the duration discrimination for the online group, which did not reach an acceptable level [$ICC_{(2,2)} = .372$].

Discussion

The current study examined the underlying constructs and test-retest reliability of auditory processing test batteries taken under laboratory and online conditions in 100 L1 and L2 speakers of English. According to a series of factor analyses, the test batteries tapped into four independent abilities—(a) pitch processing (pitch discrimination, melody reproduction), (b) spectral acuity (formant and pitch discrimination), (c) temporal acuity (duration discrimination), and (d) temporal integration (rhythm reproduction). In essence, the findings here concur with the two-factor model of auditory processing. Under this view, auditory processing is a multifaceted phenomenon consisting of different types of processing skills (acuity vs. integration) and information (spectral vs. temporal information). However, the distinction between acuity and integration abilities may be fuzzy when it comes to the processing of pitch information (pitch discrimination and melody reproduction).

There were two major findings regarding the reliability of the auditory processing test batteries. First, the ICCs could be considered as "fair" to "good" across different tasks ($ICC_{(2,2)} = .4–.6$ for discrimination; $ICC_{(2,2)} = .7–.8$ for reproduction). Second, using overall scores to calculate auditory processing appeared to be slightly more

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

reliable than subcomponent scores (ICC $_{(2,2)}$ = .625 for discrimination; ICC $_{(2,2)}$ = .843 for reproduction). All in all, it is reasonable to assume that the test batteries described here measure auditory processing ability with a fair-to-good degree of reliability, especially when the overall score is used. However, some caution may be needed if subcomponent test scores are used for interpretation (e.g., duration discrimination under online conditions).[4]

These findings lead us to a set of tentative suggestions for the measurement of auditory processing with a view of more robust, rigorous, and reliable evaluations of this construct. First, whereas researchers can use either overall or subcomponent test scores depending on their goals and hypotheses, the former may be more reliable, and may better correlate with the outcomes of L2 learning (e.g., Saito, Sun et al., 2022). Though less reliable, subcomponent test scores are important for understanding effects on language features using specific dimensions of sound (e.g., Lengeris & Hazan, 2010 for L2 vowel acquisition).

Second, temporal acuity needs to be examined with much caution because the reliability of the corresponding task (duration discrimination) was relatively low (especially under online conditions). One solution could be to use the amplitude envelope rise time discrimination task. Although this task was not included and examined in the current study, amplitude rise time discrimination has been used as an index of temporal processing in L1 acquisition literature (Goswami, 2015) and thus introduced as a substitute measure to compensate for the difficulties inherent to perceiving subtle differences in time information (i.e., duration discrimination). In this task, participants are asked to discriminate between stimuli which differ in timing of a linear ramp (e.g., 15 to 300 ms). Not only has the task been found to be reliable in our pilot project with 30 participants (e.g., Saito et al., 2020 for r = .798), but also predictive of L1 acquisition (Kalashnikova et al., 2019). Using both duration and rise time discrimination tasks complementarily may help researchers capture participants' sensitivity to temporal information.

Finally, to increase the reliability of the auditory discrimination threshold measurements, researchers are advised to ask participants to do the auditory processing tests twice and use the averaged (instead of onetime) scores to increase the reliability of the tests. In the context of L2 speech learning, there is some indication that strong associations between auditory processing and L2 proficiency are observed when participants' auditory processing abilities are measured twice (Saito, Kachlicka, Suzukida, Mora-Plaza et al., 2022).

Offline version of auditory processing test batteries

The test materials used in the current study (laboratory conditions and online platforms) have been coded in JavaScript/HTML. The test batteries with a brief user manual are currently deposited in *Tools for Second Language Speech Research and Teaching* (Mora-Plaza et al., 2022; http://sla-speech-tools.com/) and Supporting Information-A. By downloading the offline auditory processing tests onto their own computers, researchers, graduate students, and practitioners can easily assess their participants' auditory processing profiles. Following the test procedures in the "Method" section,

[4]Here, we acknowledge that the ICCs were relatively lower for the online conditions than the laboratory conditions. A reviewer pointed out that adding practice trials with feedback may help further improve the reliability of the tests. If they pass the practice part, that would be better evidence that they can complete the tests on their own.

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

participants engage in a total of six subtests that include (a) four discrimination tasks (formant, pitch, duration, and amplitude rise time) and (b) two reproduction tasks (melody, rhythm; approximately 3–5 minutes per task). Screenshots of the tasks were summarized in Figure 4.

Note that the amplitude rise time discrimination task was not included in the test-retest study presented in the preceding text. However, the test-retest reliability was relatively high in our pilot project ($r = .798$; Saito et al., 2020b) and the same procedure in the format, pitch, and duration discrimination tasks can be used.

As for the discrimination tasks, participants' test scores are recorded on a 100-point scale, with smaller scores indicating more precise acuity to particular sound dimensions. If a participant obtains 10 out of 100 points for formant (1 point/step = 2 Hz), pitch (1 point/step = 0.3 Hz), duration (1 point/step = 2.5 ms), and amplitude rise time (1 point/step = 2.8 ms), this indicate that the minimum difference they can hear is 20 Hz in F2, 3 Hz in F0, 25 ms in duration, and 28 ms in the timing of a linear ramp. As for the reproduction tasks, participants' test scores are recorded between 0–100% as an index of how precisely they can replicate the melodic and rhythmic sequences that they have heard. To calculate overall discrimination and reproduction scores, we suggest that subtest scores be first standardized, and then averaged. This is because a difference in one point/percentage entails different amounts of perceptual difficulty for both the discrimination and reproduction tasks.

Practical implications

As in L1 acquisition, the rate and ultimate attainment of adult L2 speech learning could be tied not only to experience-related factors but also to auditory processing. Recently, there is some suggestion (a) that L2 learners with different auditory processing profiles can differently benefit from different types of L2 speech training; and (b) that auditory processing scores can be used to diagnose profile-matched training with a view of optimal L2 speech learning. For example, acuity can be associated with gains when learners engage in perception-based training (e.g., Lengeris & Hazan, 2010 for high variability phonetic training); and integration can be predictive of those who can show more improvement resulting from production-based training (e.g., Li & DeKeyser, 2017 for repetition training; Shao et al., 2022 for shadowing). More importantly, it has been claimed that the quality of instructed L2 speech learning could be limited among certain learners with lower auditory processing scores (Perrachione et al., 2011; Ruan & Saito, 2022). Thus, scholars have begun to examine the extent to which provision of *auditory* training prior to phonetic training can help all learners *equally* achieve successful L2 speech learning (Saito, Petrova, et al., 2022). For a comprehensive overview on the pedagogical potential of treatment-aptitude interaction in instructed L2 speech learning, see Saito (ibid.).

Future directions

To close, we would like to provide a range of future directions that researchers can pursue to unravel the relationship between auditory processing and L2 speech learning. First and foremost, though tentative, the results of post-hoc analyses showed that the reliability of the tests was comparable whether the tests were conducted under laboratory versus online conditions. The lack of substantial influence of testing condition (laboratory vs. online) indicates that the instructions and modalities of the tasks

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

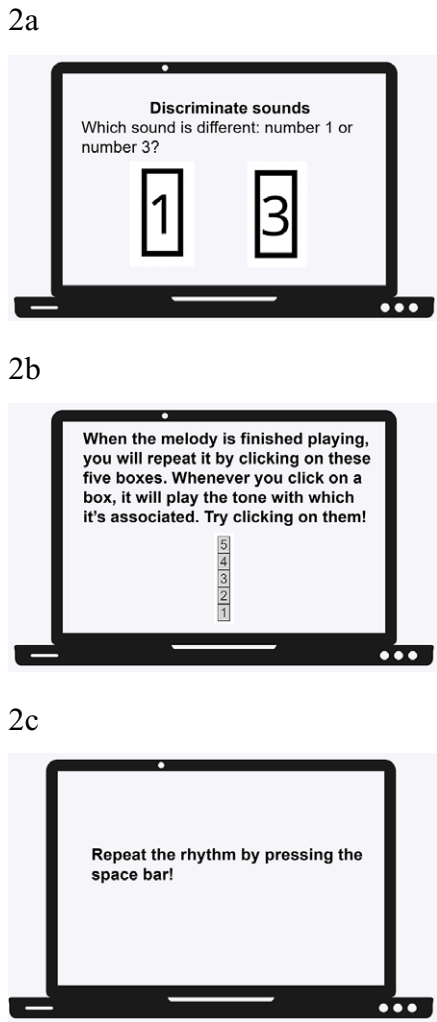

Figure 4. Screenshots of offline auditory processing test batteries: Formant discrimination (2a), melody reproduction (2b) and rhythm reproduction (2c).

(discrimination, reproduction) may be adequately intuitive such that participants can complete the tests on their own. However, we acknowledge that the post-hoc analyses of the test-retest reliability under two different testing conditions (lab vs. online) need to be interpreted with much caution. The current study was not carefully designed to answer this methodological question. In fact, the two types of data were different in terms of not only testing conditions (lab vs. online) but also a range of biographical and linguistic factors (e.g., classroom vs. immersion experience; L1 vs. L2 English speakers). We make a strong call for future studies to revisit the topic (the reliability of lab- vs. online-based auditory processing tests) with carefully controlled research design wherein the same participants with homogeneous backgrounds take the tests multiple times in lab and online settings.

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

Next, we acknowledge that the current study relied on product-oriented evidence (i.e., participants' scores) without taking into consideration any underlying processes (i.e., participants' inner perception of the tests). Future studies should ask participants to verbalize what they were thinking while completing the task. This line of work could help reveal the extent to which the participant's response was random or not.

Existing studies suggest that different components of auditory processing (acuity vs. integration of spectral vs. temporal information) can be uniquely associated with different aspects of L2 speech learning (phonology vs. lexicogrammar, naturalistic vs. classroom, perception- vs. production-based practice). Notably, most of the existing evidence is cross-sectional in nature. Thus, we strongly call for more intervention studies that examine the mediating roles of auditory processing in L2 speech learning from a longitudinal perspective. Such studies would allow researchers to track how participants with diverse auditory processing profiles can enhance the segmental, prosodic, temporal, lexical, and morphosyntactic dimensions of their L2 speech under different learning conditions (e.g., focus on meaning vs. form instruction; Spada & Tomita, 2010; perception- vs. production-based practice; Shintani et al., 2013).

Furthermore, more work is needed to further examine the construct validity of the auditory processing tests. Traditionally, construct validation is to interpret a test as "a measure of some attribute or quality which is not 'operationally defined'" (Cronbach & Meehl, 1955, 282). To date, it has remained controversial the extent to which the task formats that scholars have adopted (discrimination and reproduction) tap into what they claim to measure (the acuity and integration aspects of auditory processing, respectively). By definition, auditory processing is essentially a bottom-up and implicit phenomenon. Thus, the use of behavioral tasks as the outcome measures of auditory processing has been questioned as they inevitably involve conscious, attentional, and cognitive processing of sounds (e.g., Clinard et al., 2010; Snowling et al., 2018). One promising direction is to investigate the unique contribution of auditory processing to L2 speech learning when other cognitive skills are controlled for statistically. In the context of adult L2 learners, for example, Saito, Cui et al. (2022) recently provided cross-sectional evidence that perceptual acuity (assessed using the discrimination tasks) and explicit analytic abilities (assessed using the simple and complex working memory tasks) *independently* predicted variance in L2 speech acquisition. Following this line of thought, it would be intriguing to examine how participants with different levels of perceptual-cognitive abilities (including auditory processing) develop their L2 speech abilities when undergoing different types of training (e.g., input- vs. output-based; form- vs. meaning-oriented). Such studies could help shed light on how auditory processing and other cognitive abilities interact to affect L2 speech learning under different input and output conditions.

Scholars have examined the potential of neural measures as a valid way to assess the state of auditory processing *without* the influence of cognitive state (i.e., awareness and attention; for an overview on *implicit* auditory processing, see Sun et al., 2021). One such example can be found in an electroencephalographic (EEG) paradigm known as the frequency following response (FFR). The FFR can be used as an index of neural encoding to the frequency content of periodic auditory stimuli. Because the FFR is very rapidly generated (only 10 ms after a stimulus is presented to a participant) and requires phase-locking of neurons to high-frequency components of sounds, the neural generators of the responses are mainly located in the auditory brainstem (White-Schwoch et al., 2017), with an additional cortical contribution for lower frequencies of the FFR [< 150Hz], see Coffey et al., 2016). The test-retest reliability and variability of FFR is likely robust (Easwar et al., 2020; Hornickel et al., 2012; Sun et al., 2021 for

https://doi.org/10.1017/S027226312200047X Published online by Cambridge University Press

between-session correlations of $r > 0.8$). The FFR has been found to relate to speech perception in noise (e.g., Anderson et al., 2010), phonological awareness (e.g., Banai et al., 2009), and global reading abilities (e.g., Hornickel & Kraus, 2013). More recently, scholars have begun to link FFR to a range of L2 speech abilities (e.g., Omote et al., 2017 for speech perception; Saito et al., 2019; Kachlicka et al., 2019 for morphosyntax). Notably, some studies have shown that FFR performance can differ from the outcomes of AXB discrimination tasks, suggesting that the neurophysiological and behavioral tasks may tap into different (implicit vs. explicit) constructs of auditory processing (Clinard et al., 2010). Given that scholars have argued that the attainment of high-level L2 proficiency requires both explicit and implicit aptitude (Doughty, 2019), future studies should further explore how explicit auditory processing (measured using discrimination and reproduction tasks) is related to implicit auditory processing (measured through FFR), and whether such abilities interact to determine the outcomes of naturalistic and instructed L2 speech learning.

Finally, it has been shown that auditory processing (measured using the discrimination and reproduction tasks) is replicable and stable over time ($ICC_{(2,2)} = .4–.8$) as well as predictive of various aspects of L2 speech learning. To obtain a full-fledged understanding of how auditory processing facilitates L2 speech learning, we still need to know precisely what underlines the development, attainment, and maintenance of auditory processing abilities, especially in adulthood. Whereas auditory processing is subject to age-related decline (e.g., Schneider et al., 2002), it can be enhanced using extensive amounts of music and bilingual experience (Krizman et al., 2015), as well as by as little as a few hours of training (Micheyl et al., 2006). In terms of adult L2 speakers, some research has demonstrated that the acuity component of auditory processing may be linked to age-related factors (chronological age, age of acquisition) but not to experience-related factors (length of residence, the frequency of L2 use; e.g., Saito, Sun et al., 2022). However, reproduction could be linked to the extent to which L2 learners have practiced a target language in foreign language classroom learning (Saito, Suzukida et al., 2021). More research is needed to clarify why certain adult individuals possess high-level auditory processing abilities (exerting positive influences on L2 speech learning), and why others may fail to meet certain thresholds (inhibiting L2 speech learning). This latter line of research is particularly crucial for (a) identifying the source(s) of L2 learning difficulty, and (b) understanding how and whether the provision of intensive auditory processing training can help learners equally understand, speak, and master a target language regardless of initial auditory processing ability (however, see McArthur et al., 2008 for the amenability of auditory processing).

Acknowledgments. We are grateful to the following team members for their great contributions to every stage of our collaboration projects between 2018 and 2022: Hui Sun, Magdalena Kachlicka, Yui Suzukida, Ingrid Mora-Plaza, Katya Petrova, Ruan Yaoyao, Oscar Macmillan, Sascha Kroeger, Kotaro Takizawa, Haining Cui, Diego Elisandro Dardon, Mai Tran, and Viktoria Magne. We thank Joan Mora and Josh Frank with their assistance with data collection and analyses. The work presented here was funded by Leverhulme Trust (RPG-2019-039), Spencer Foundation (202100074), and Economic and Social Research Council (ES/V007955/1).

Data availability statement. The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://osf.io/bkj6v/?.

Competing interests. The authors declare none.

References

Anderson, S., Skoe, E., Chandrasekaran, B., Zecker, S., & Kraus, N. (2010). Brainstem correlates of speech-in-noise perception in children. *Hearing Research*, *270*, 151–157. https://doi.org/10.1016/j.heares.2010.08.001

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://link.springer.com/article/10.3758/s13428-019-01237-x

Baker, A. (2014). Exploring teachers' knowledge of second language pronunciation techniques: Teacher cognitions, observed classroom practices, and student perceptions. *Tesol Quarterly*, *48*, 136–163. https://doi.org/10.1002/tesq.99

Banai, K., Hornickel, J., Skoe, E., Nicol, T., Zecker, S., & Kraus, N. (2009). Reading and subcortical auditory function. *Cerebral Cortex*, *19*, 2699–2707. https://doi.org/10.1093/cercor/bhp024

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

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284. https://doi.org/10.1037/1040-3590.6.4.284

Clinard, C. G., Tremblay, K. L., & Krishnan, A. R. (2010). Aging alters the perception and physiological representation of frequency: Evidence from human frequency-following response recordings. *Hearing Research*, *264*, 48–55. https://doi.org/10.1016/j.heares.2009.11.010

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*, 1–11. https://www.nature.com/articles/ncomms11070

Corriveau, K. H., & Goswami, U. (2009). Rhythmic motor entrainment in children with speech and language impairments: Tapping to the beat. *Cortex*, *45*, 119–130. https://doi.org/10.1016/j.cortex.2007.09.008

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281. https://doi.org/10.1037/h0040957

Dellatolas, G., Watier, L., Le Normand, M. T., Lubart, T., & Chevrie-Muller, C. (2009). Rhythm reproduction in kindergarten, reading performance at second grade, and developmental dyslexia theories. *Archives of Clinical Neuropsychology*, *24*, 555–563. https://doi.org/10.1093/arclin/acp044

Doughty, C. J. (2019). Cognitive language aptitude. *Language Learning*, *69*, 101–126. https://doi.org/10.1111/lang.12322

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*, 150–164. https://doi.org/10.1097/AUD.0000000000000739

Field, A. (2009). *Discovering statistics using SPSS*. 3rd ed. SAGE Publications.

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. https://doi.org/10.3389/fnhum.2014.00392

Goswami, U. (2015). Sensory theories of developmental dyslexia: Three challenges for research. *Nature Reviews Neuroscience*, *16*, 43–54. https://www.nature.com/articles/nrn3836

Grahn, J. A., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience*, *19*, 893–906. https://doi.org/10.1162/jocn.2007.19.5.893

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.) Prentice-Hall International, Inc.

Hornickel, J., Knowles, E., & Kraus, N. (2012). Test-retest consistency of speech-evoked auditory brainstem responses in typically-developing children. *Hearing Research*, *284*, 52–58. https://doi.org/10.1016/j.heares.2011.12.005

Hornickel, J., & Kraus, N. (2013). Unstable representation of sound: A biological marker of dyslexia. *Journal of Neuroscience*, *33*, 3500–3504. https://doi.org/10.1523/JNEUROSCI.4205-12.2013

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. https://doi.org/10.1016/j.bandl.2019.02.004

Kalashnikova, M., Goswami, U., & Burnham, D. (2019). Sensitivity to amplitude envelope rise time in infancy and vocabulary development at 3 years: A significant relationship. *Developmental Science*, *22*, e12836. https://doi.org/10.1111/desc.12836

Kempe, V., Bublitz, D., & Brooks, P. J. (2015). Musical ability and non-native speech-sound processing are linked through sensitivity to pitch and spectral information. *British Journal of Psychology*, *106*, 349–366. https://doi.org/10.1111/bjop.12092

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*, 418–435. https://doi.org/10.1121/1.2743154

Krizman, J., Slater, J., Skoe, E., Marian, V., & Kraus, N. (2015). Neural processing of speech in children is influenced by extent of bilingual experience. *Neuroscience Letters*, *585*, 48–53. https://doi.org/10.1016/j.neulet.2014.11.011

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*, 1279–1292. https://link.springer.com/article/10.3758/BF03194543

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*, 3757–3768. https://doi.org/10.1121/1.3506351

Levitt, H. C. C. H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*, 467–477. https://doi.org/10.1121/1.1912375

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*, 593–620.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PloS One*, *14*, e0219854.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and Speech*, *29*, 3–11. https://doi.org/10.1177/002383098602900102

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed), *Advancing quantitative methods in second language research* (pp. 182–212). Routledge.

McArthur, G. M., Ellis, D., Atkinson, C. M., & Coltheart, M. (2008). Auditory processing deficits in children with reading and language impairments: Can they (and should they) be treated? *Cognition*, *107*, 946–977. https://doi.org/10.1016/j.cognition.2007.12.005

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*, 36–47. https://doi.org/10.1016/j.heares.2006.05.004

Moore, B. C. (2012). *An Introduction to the Psychology of Hearing*. Brill.

Mora-Plaza, I., Saito, K., Suzukida, Y., Dewaele, J-M., & Tierney, A. (2022). *Tools for second language speech research and teaching*. http://sla-speech-tools.com. http://doi.org/10.17616/R31NJNAX

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*, 15953–15958. https://doi.org/10.1073/pnas.1204319109

Omote, A., Jasmin, K., & Tierney, A. (2017). Successful non-native speech perception is linked to frequency following response phase consistency. *Cortex*, *93*, 146–154. https://doi.org/10.1016/j.cortex.2017.05.005

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in Systems Neuroscience*, *8*, 57. https://doi.org/10.3389/fnsys.2014.00057

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*, 461–472. https://doi.org/10.1121/1.3593366

Povel, D., & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*, 411–440.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*, 7865–7870. https://doi.org/10.1073/pnas.0509989103

Qin, Z., Zhang, C., & Wang, W. S. Y. (2021). The effect of Mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones. *The Journal of the Acoustical Society of America*, *149*, 435–446. https://doi.org/10.1121/10.0003330

Ramus, F. (2003). Developmental dyslexia: Specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*, *13*, 212–218. https://doi.org/10.1016/S0959-4388(03)00035-7

Rosen, S. (2003). Auditory processing in dyslexia and specific language impairment: Is there a deficit? What is its nature? Does it explain anything? *Journal of Phonetics*, *31*, 509–527. https://doi.org/10.1016/S0095-4470(03)00046-9

Ruan, Y., & Saito, K. (2022). *Communicative focus on phonetic form revisited: Domain-general auditory difficulties limit instructed second language speech learning* [Manuscript submitted for publication].

Saito, K. (in press). How does having a good ear promote successful second language speech acquisition in adulthood? Introducing auditory precision hypothesis-L2. *Language Teaching*.

Saito, K., Cui, H., Suzukida, Y., Dardon, D. E., Suzuki, Y., Jeong, H., … & Tierney, A. (2022). Does domain-general auditory processing uniquely explain the outcomes of second language speech acquisition, even once cognitive and demographic variables are accounted for? *Bilingualism: Language and Cognition*, *25*, 856–868. https://doi.org/10.1017/S1366728922000153

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing, age, experience, and post-pubertal L2 speech learning: A behavioral and neurophysiological investigation. *Journal of Memory and Language*, *115*, 104–168. https://doi.org/10.1016/j.jml.2020.104168

Saito, K., Kachlicka, M., Suzukida, Y., Mora-Plaza, I., Ruan, Y., & Tierney, A. (2022). *Auditory processing as composite abilities underlying successful second language acquisition: Acuity, attention, and integration model* [Manuscript in preparation].

Saito, K., Kachlicka, M., Suzukida, Y., Petrova, K., Lee, B. J., & Tierney, A. (2022). Auditory precision hypothesis-L2: Dimension-specific relationships between auditory processing and second language segmental learning. *Cognition*, *229*, 105236. https://doi.org/10.1016/j.cognition.2022.105236

Saito, K., Petrova, K., Suzukida, Y., Kachlicka, M., & Tierney, A. (2022). Training auditory processing promotes second language speech acquisition. *Journal of Experimental Psychology: Human Perception and Performance*, *48*, 1410–1426. https://doi.org/10.1037/xhp0001042

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652–708. https://doi.org/10.1111/lang.12345

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*, 1123–1140. https://doi.org/10.1017/S1366728918000895

Saito, K., Sun, H., & Tierney, A. (2020a). Domain-general auditory processing determines success in second language pronunciation learning in adulthood: A longitudinal study. *Applied Psycholinguistics*, *41*, 1083–1112. https://doi.org/10.1017/S0142716420000491

Saito, K., Sun, H., & Tierney, A. (2020b). Brief report: Test-retest reliability of explicit auditory processing measures. *BioRxiv*. https://doi.org/10.1101/2020.06.12.149484

Saito, K., Sun, H., Kachlicka, M., Alayo, J. R. C., Nakata, T., & Tierney, A. (2022). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*, *44*, 57–86. https://doi.org/10.1017/S0272263120000467

Saito, K., Suzukida, Y., Tran, M., & Tierney, A. (2021). Domain-general auditory processing partially explains second language speech learning in classroom settings: A review and generalization study. *Language Learning*, *71*, 669–715. https://doi.org/10.1111/lang.12447

Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *56*, 139. https://doi.org/10.1037/h0087392

Shao, Y., Saito, K., & Tierney, A. (2022). How does having a good ear promote instructed second language pronunciation development? Roles of domain-general auditory processing in choral repetition training. *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.3120

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*, 296–329. https://doi.org/10.1111/lang.12001

Smith, J. (2007). *Introduction to digital filters with audio applications*. W3K Publishing.

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*, 1270–1282. https://doi.org/10.1177/0956797618763090

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*, 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Sun, H., Saito, K., & Tierney, A. (2021). A longitudinal investigation of explicit and implicit auditory processing in L2 segmental and suprasegmental acquisition. *Studies in Second Language Acquisition*, *43*, 551–573. https://doi.org/10.1017/S0272263120000649

Surányi, Z., Csépe, V., Richardson, U., Thomson, J. M., Honbolygó, F., & Goswami, U. (2009). Sensitivity to rhythmic parameters in dyslexic children: A comparison of Hungarian and English. *Reading and Writing*, *22*, 41–56. https://link.springer.com/article/10.1007/s11145-007-9102-x

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America*, *110*, 2085–2095. https://doi.org/10.1121/1.1404973

Tallal, P. (2004). Improving language and literacy is a matter of time. *Nature Reviews Neuroscience*, *5*, 721–728. https://www.nature.com/articles/nrn1499

Thompson, E. C., White-Schwoch, T., Tierney, A., & Kraus, N. (2015). Beat synchronization across the lifespan: Intersection of development and musical experience. *PLoS One*, *10*, e0128839. https://doi.org/10.1371/journal.pone.0128839

Tierney, A., & Kraus, N. (2014). Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (PATH). *Frontiers in Human Neuroscience*, *8*, 949. https://doi.org/10.3389/fnhum.2014.00949

Tierney, A., White-Schwoch, T., MacLean, J., & Kraus, N. (2017). Individual differences in rhythm skills: Links with neural consistency and linguistic ability. *Journal of Cognitive Neuroscience*, *29*, 855–868. https://doi.org/10.1162/jocn_a_01092

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*, 5095–5115.

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

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M., & Merzenich, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature*, *387*, 176–178. https://www.nature.com/articles/387176a0

Cite this article: Saito, K. and Tierney, A. (2022). Domain-general auditory processing as a conceptual and measurement framework for second language speech learning aptitude: A test-retest reliability study. *Studies in Second Language Acquisition,* 1–25. https://doi.org/10.1017/S027226312200047X