OXFORD

Cerebral Cortex, 2025, **35**, bhaf072
https://doi.org/10.1093/cercor/bhaf072
Advance access publication date 2 April 2025
Original Article

Brain activation patterns reflecting differences in music training: listening by ear vs. reading sheet music for the recognition of contexts and structures in a composition

Reiya Horisawa[1], Keita Umejima[1], Seizo Azuma[2,3], Takeaki Miyamae[3,4], Ryugo Hayano[3], and Kuniyoshi L. Sakai[1,*]

[1]Department of Basic Science, Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
[2]Department of Instrumental Music, Faculty of Music, Tokyo University of the Arts, 12-8 Ueno Park, Taito-ku, Tokyo 110-8714, Japan
[3]Suzuki School of Music, The Talent Education Research Institute, 3-10-3 Fukashi, Matsumoto-shi, Nagano 390-8511, Japan
[4]Department of Psychiatry, University of Pittsburgh School of Medicine, Thomas Detre Hall, 3811 O'Hara Street, Pittsburgh, PA 15213, United States

*Corresponding author: Kuniyoshi L. Sakai, Department of Basic Science, Graduate School of Arts and Sciences, The University of Tokyo, Komaba, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan. Email: sakai@sakai-lab.jp

When practicing a new piece of music, what are the neural substrates influenced by short-term training such as listening to recorded sources or reading sheet music? Do those neural mechanisms reflect the effects of long-term training in music? In the present functional magnetic resonance imaging study with intermediate piano players in the middle of acquiring advanced knowledge and skills in music, we compared short-term training of listening to recorded pieces ("Listen") and reading sheet music ("Read"). Participants were "Multi-" and "Mono-instrumentalist" groups according to whether they played multiple instruments or only the piano. We used an error-detection task with music stimuli including *structural* errors made by swapping 2 phrases within a composition, thereby focusing on contextual comprehension of musical phrases. Overall performances were significantly better under Listen than under Read, and significantly better in Multi than in Mono. Moreover, we observed left-lateralized frontal activations under Listen for Multi, whereas bilateral temporo-frontal regions were activated under Read for both groups. Focusing on individual differences under Read, we found a positive correlation between the frontal activations and the accuracy rates for Mono. Overall, our results elucidate how the neural substrates of judgments on structures and context in music are influenced by both long-term and short-term training.

Keywords: auditory areas; context; music; left frontal regions; training.

Introduction

In music acquisition, the generally assumed superiority of music sounds over sheet music has not been conclusively confirmed. Sheet music can provide connections/breaks (e.g. slurs and breath marks) and expression markings (e.g. *p*, *f*, *cresc.*, and *marcato*) that provide information on articulation and related phrasal structures, although such readings are open to individuals' interpretations in music (Bernstein 1976; Kramer 2010). Thus the retrieval of vocal or instrumental sounds from sheet music, as well as the reconstruction of phrasal structures, is reliant on the individual player's experiences and understanding of music styles (Dart 1954). In contrast, exposure to music sounds provides, in a direct manner, the structural information that facilitates contextual interpretation (including phrasing and articulation), eliminating the first retrieval processes. It has been proposed that phrasal structures in tonal music are analyzed in the same way as phrase structures in natural language (Lerdahl and Jackendoff 1983; Jackendoff and Lerdahl 2006). On the other hand, there is a clear difference between what is comprehensible in music and language; music has unique roles in communication and social bonding different from those in language (Cross 2014; Savage et al. 2021). Here, we focused on the contextual and structural aspects of composition and hypothesized that brain functions measured by cortical activations are differentially influenced by the modalities of music training, i.e. training using sheet music or more *naturally*, training through listening by ear.

In our previous study on music using an error-detection task and functional magnetic resonance imaging (fMRI), we examined the long-term effects of training by comparing 3 groups of students varying in age of acquisition (AOA) and/or methods for lessons in music (Sakai et al. 2022). One of the tested groups was Suzuki students who had practiced the violin. The Suzuki Method is a series of music education courses inspired by the natural mode of acquisition of a mother tongue (Suzuki 2013) and prioritizes regularly listening to recorded performances of virtuosi and outstanding musicians before being introduced to notation reading and theory (Steinschaden and Zehetmair 1985). Group differences were clearly observed in brain activation patterns under each of the 4 tested music conditions (pitch, tempo, stress, and articulation). The errors in the articulation condition included "staccato" instead of "legato," "decrescendo" instead of "crescendo," and monotonously without intonation, all of which spanned multiple notes and thus affected phrasal structures. Under the articulation condition, activations in the left lateral premotor cortex (L. LPMC) and left opercular/triangular parts of the inferior frontal gyrus (L. F3op/F3t) were observed in all 3 groups. These regions have been identified as "grammar centers"

Received: December 13, 2024. **Revised:** January 30, 2025. **Accepted:** February 27, 2025

related to syntactic structures in both first language (L1) and second language (L2) (Sakai 2005) and are the regions thought to be involved also in analyzing phrasal structures in music, as observed under the articulation condition. As such, they were expected to be involved under the conditions in our present study as well.

In the present study, we recruited intermediate piano players in the middle of acquiring advanced knowledge and skills in music, including Suzuki students. We focused on listening to recorded pieces on a compact disk (CD) ("Listen" training condition) and reading sheet music ("Read" training condition), both of which represented *short-term* training effects (Fig. 1A). These methods were typical music training strategies for the learning of unfamiliar pieces of music, especially for intermediate players. When learning a new piece to play, it is common for amateur musicians to listen to recorded versions of new pieces to gain familiarity with tempo indications, phrasings, etc., but it would be uncommon to read the sheet music without singing or playing the piece at all. In the present study, we examined the conditions of listening or reading separately, but we set 2 d for playing the piano while reading sheet music. In the error-detection task, we auditorily presented music stimuli including *structural* errors, which were made by swapping 2 phrases (normal: Fig. 1B; with errors: Fig. 1C). Note that the swapped phrases were normal in themselves, but they were regarded as unnatural based on the *context* of preceding or following phrases. Therefore, the acquisition of contextual knowledge was necessary to detect those errors, such that memory factors or detailed familiarization with the stimuli did not impact the knowledge acquisition targeted in this study. In general, the recognition of such contextual information is a prerequisite for predictive coding and planning ahead to play an instrument. In accordance with the generally held hypothesis regarding the 2 main modalities of music training, i.e. training using sheet music or training through listening by ear, we predicted that the Listen training condition would achieve better task performances than the Read training condition. We further focused on the *long-term* effects of music training by splitting participants into 2 instrumentalist groups: a group of participants who played multiple instruments ("Multi" group) and a group that played only the piano ("Mono" group). We conjectured that the abstract phrasal structures would be better acquired by the Multi group, because the benefits of playing multiple instruments would extend beyond the acquisition of techniques for individual instruments to the enhancement of cumulative effects in music interpretation. These groups roughly correspond to *multilinguals* (L1, L2, . . .) and *monolinguals* (L1 alone). In our previous study on multilingualism, we found that the bilateral LPMC, F3op/F3t/F3O (including the orbital part of the inferior frontal gyrus), and superior/middle temporal gyri (STG/MTG) were mainly involved (Umejima et al. 2021). Moreover, we suggested the possibility that "[c]ortical activations increase initially at the onset of acquisition, followed by the maintenance of the activations and then a fall in activations during consolidation of linguistic competence" (Sakai 2005). Such multiphase changes of activations may apply to music acquisition as well, reflecting differences in long-term music training. It is thus interesting to see the short-term and long-term effects of music acquisition in terms of brain activations.

Materials and methods

Participants

We recruited a total of 45 intermediate piano players, who were deemed to have sufficient skills to perform 2 piano pieces: *Etudes de Mecanisme* (*Thirty New Studies In Technics*) by Carl Czerny (Op. 849, composed in 1856) or *Sonatinen Album* compiled by Louis Köhler and Adolf Ruthardt (Peters Edition, published in 1897). We focused on piano players, who were used to simultaneously performing at least 2 separate parts (right and left hands). The participants consisted of 30 students (mostly secondary-school students) taking lessons with the Suzuki Method in the Tokyo metropolitan area and 15 college students and adults recruited through a website, who either took lessons with other piano-training methods (13 participants), took lessons with the Suzuki Method briefly in childhood (1 participant, not regarded as a Suzuki student here), or were completely self-taught (1 participant). We excluded 2 Suzuki students who took medications and 1 Suzuki student who quit the experiment before obtaining structural MRI data. All remaining participants were right-handed, according to positive laterality quotients of handedness (LQ) tested by the Edinburgh Handedness Inventory (Oldfield 1971), and had no history of neurological disorders.

Four volunteers among the Suzuki students (2 females and 2 males, 15.9 ± 1.9 yr old [mean ± SD]) were assigned to a *reference* ("Ref") group from which only behavioral data were collected; these volunteers wore normal headphones in a quiet environment and did not undergo MR scanning.

The remaining 38 participants answered a questionnaire regarding their musical (instrument and/or vocal) training, including both private/group lessons and self-education. Nineteen participants (including 11 Suzuki students) had experience in playing 1 or 2 musical instruments other than the piano [French horn: 4 participants; trumpet: 3; violin: 3; and other: 11] for more than 1 yr, and were thus designated as the *Multi-instrumentalist* ("Multi") group (Table 1). The remaining 19 participants (including 12 Suzuki students) played only the piano and were designated as the *Mono-instrumentalist* ("Mono") group, which was age-matched with the Multi group ($t[36] = 0.4$, $P = 0.7$). For each instrument, the period between her/his AOA and the latest age with training was defined as the duration of exposure (DOE). The DOE was measured separately for the piano and the other instruments (see Table 1); when 2 other instruments were learned simultaneously, the overlapped period was not duplicated for the DOE. If there was an absence from practice of more than 6 months, the period of absence was subtracted to obtain the DOE. Musical training as a part of the school curriculum was not included in the DOE, because it consisted of only 40 h of training each year and thus had little effect on the age-matched groups, except for 2 members of the Multi group who were in a secondary school (80 h each year) and a college (180 h each year) specializing in music, respectively. Moreover, the total time spent practicing instruments was also estimated for each participant by adding together the number of hours of lessons and practices. Regarding piano playing, the AOA, DOE, and practicing hours were not normally distributed in at least either group (Shapiro–Wilk tests, $P < 0.05$). We thus used nonparametric tests for group comparisons, and these values were comparable between the groups (2-sample Kolmogorov–Smirnov tests; AOA: $P = 0.05$; DOE: $P = 0.04$; practicing hours: $P = 0.1$; Bonferroni corrected $\alpha = 0.017$; see Table 1). The practice hours on the other instrument(s) in the Multi group were also comparable to those on the piano in the Mono group ($P = 0.5$).

All of the participants, as well as their legal guardians for those younger than 18 yr of age, provided their written informed consent to participate in this study after the nature and possible consequences of the study were explained. Approval for these experiments was obtained from the institutional review board of the University of Tokyo, Komaba Campus (approval nos. 497–6,

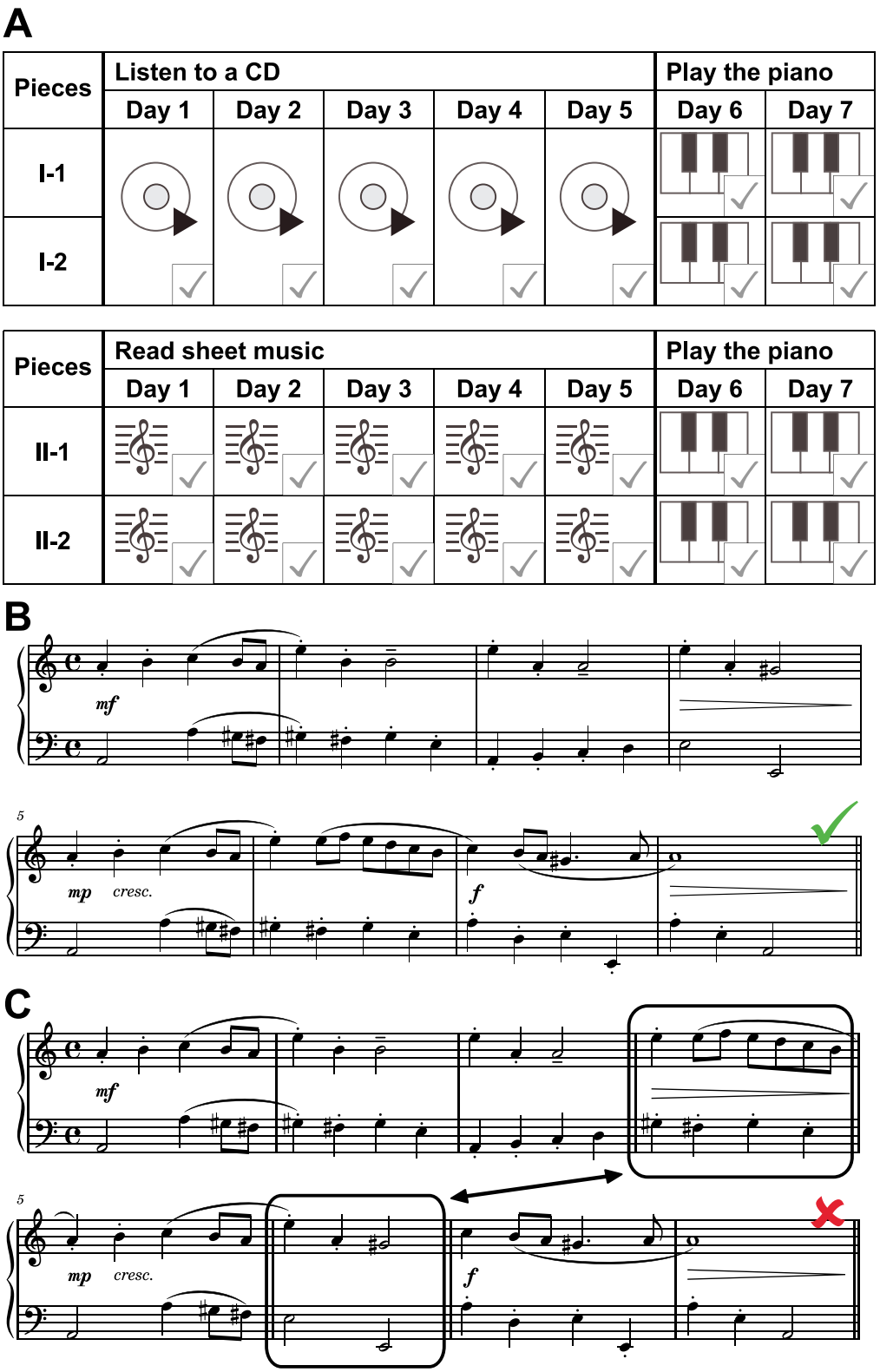

Fig. 1. Protocol for examining contextual comprehension of music associated with training. (A) An example of training for 7 d on 4 music pieces, grouped into Sets I (pieces I-1 and I-2) and II (pieces II-1 and II-2). During the first 5 d, the participants listened to recorded pieces on a CD for one Set (Set I in this case), and read sheet music for the other Set. These were designated the "Listen" and "Read" training conditions, respectively. During the last 2 d, the participants trained by playing the piano while reading the sheet music for both Sets. Participants placed a check mark in the bottom-right corner of each box to self-report her/his fulfillment of training. (B) An example of a *normal* stimulus (✓), which was always auditorily presented in the scanner. In an error-detection task under the "Context" condition, participants listened to an excerpt of recorded pieces, and judged whether there was an unnatural portion in the excerpt. The initial section from piece II-2 (*Entrée* in A minor) is shown. (C) The *unnatural* stimulus (✗) formed from the normal one shown in (B). Two phrases surrounded by 2 boxes were swapped (denoted by a double-headed arrow) to construct an unnatural stimulus. Swapped phrases preserved the major rules of counterpoint, as well as natural flows in harmony, but produced *structural* changes. Note that the swapped phrases were normal in themselves, but they were regarded as errors in the task based on the *context* of preceding and following phrases. These auditory stimuli were presented for 18 s, including a few more bars with fade-out. Supplementary_Material_1 (Fig. 1B) and Supplementary_Material_2 (Fig. 1C) are provided for full-length auditory stimuli.

Table 1. Participant characteristics in the Multi- and Mono-instrumentalist groups.

Group	N	Age (yr)	Piano			Other instruments			LQ
			AOA (yr)	DOE (yr)	Practice (h)	AOA (yr)	DOE (yr)	Practice (h)	
Multi	19 (14 f.)	20.0 ± 7.1	3.6 ± 1.0	14 ± 5.9	3200 ± 2000	11 ± 3.9	6.2 ± 5.0	2200 ± 1800	82 ± 20
Mono	19 (14 f.)	19.0 ± 5.8	4.5 ± 1.1	10 ± 3.8	2400 ± 1700	–	–	–	90 ± 14

Data are shown as the mean ± SD. N, number of participants. f, female. AOA, age of acquisition for a musical instrument. DOE, duration of exposure. Practice, approximate total time spent practicing instrument(s). LQ, laterality quotient of handedness.

497–7, and 497–8). All research studies were performed in accordance with the Declaration of Helsinki, the Singapore Statement on Research Integrity, and the relevant guidelines/regulations in Japan (the Science Council of Japan and the Japan Society for the Promotion of Science).

Stimuli

We tested music stimuli that were mostly unfamiliar to the participants before training. These consisted of short excerpts from typical pieces of Western classical piano music, which were suitable and instructive for intermediate piano players we studied here, in that these pieces serve as the basis for understanding music structures and playing styles. We used 4 pieces, labeled I-1 and I-2 (grouped into Set I) and II-1 and II-2 (grouped into Set II), as shown below.

I-1: *Minuet* in A minor by Johann Sebastian Bach (BWV Anh. 120, composed in 1725)

I-2: *Marcia* (*King William's March*) in D major by Jeremiah Clarke (composed in 1702)

II-1: *Minuet* in G major by Georg Böhm (composed in 1725)

II-2: *Entrée* in A minor by Leopold Mozart (known as part of the *Notebook for Wolfgang*; the *Entrée* has a dedication dated 1762, but this inscription has been disputed).

We used these pieces with a 2-voice counterpoint, which contains richer structural information than monophony. The pieces I-1 and II-2 began with a theme in a minor key, which was then recapitulated in its relative major key and then played a third time in the original key. On the other hand, the pieces I-2 and II-1 began with a theme in a major key, followed by a variation in its dominant key, and finally returned to the original key. These 4 pieces were played by each of 2 professional pianists, Seizo Azuma (S.A., one of the authors) and Sakiko Ishikawa (S.I.), and digitally recorded. These stimuli were musically varied between the pianists in that the key touch and articulatory interpretation were slightly different. To assess memory factors or familiarization effects on the stimuli in detail, the stimuli used for the 5-d training with a CD were played by S.A., and the stimuli used for the task were played by S.A. and S.I.; the task performances were then compared between the stimuli played by the 2 pianists. Pieces I-1 and II-1 were played at around 120 beats per minute (bpm), whereas I-2 and II-2 were played at around 144 bpm. Throughout the experiment, we did not inform the participants about the titles or composers of the pieces.

We asked the participants to report their familiarity with each piece before the training on a 3-point scale: known, somewhat familiar, and unfamiliar. All participants were unfamiliar with all 4 pieces, except for 1 participant of the Multi group, who was somewhat familiar with pieces I-1 and II-2 but unfamiliar with the others, and 1 participant of Mono, who was somewhat familiar with II-2 but unfamiliar with the others.

For music training, all participants from both groups were asked to familiarize themselves with these 4 pieces at home. The training consisted of 7 consecutive days, and the participants self-reported their fulfillment of training every day (Fig. 1A). Seven participants (5 in Multi, 2 in Mono) had skipped training for 1 or 2 of the first 5 d, whereas all other participants fulfilled training for all 7 d. During the first 5 d of training, there were 2 conditions: the *Listen* and *Read* training conditions (see Fig. 1A). The Sets I and II were counter-balanced between Listen and Read among the participants from both groups, such that Set I was used for 19 participants under Listen, and Set II for the other 19 participants, and vice versa under Read. The participants listened to recorded pieces on a CD (5 times per piece for each day) under Listen, whereas under Read, they read sheet music for a duration equivalent to the total listening time under Listen. Among all the participants, we combined both Sets I and II for each of the training conditions to analyze behavioral and functional data.

On the last 2 of 7 training days, the participants actually played the piano (using a full-size piano or portable piano keyboard) (see Fig. 1A) to consolidate the short-term training effects for Listen and Read, and to check what was learned for these 4 pieces by themselves. They played the piano while reading sheet music for a duration equivalent to the total listening time under the Listen condition. We conducted MR scans on the next day, and compared any resultant changes due to training under the Listen and Read conditions. Participants in the Ref group were not engaged in the 7-d training; they thus served as a reference group for task performances without short-term training.

In this study, we newly prepared "unnatural" stimuli, which consisted of the same short excerpts but included 2 swapped phrases of the same length, 1 or 2 bars each. Each swapped phrase was so short that the overall composition was retained. The swapped positions of the phrases were not random, but chosen so as to maintain the major rules of counterpoint (Kennan 1998), without producing unnatural flows in harmony or excessive leaps in pitch. The number of bars separating the 2 phrases ranged from 0 to 18. For example, in the case of *Entrée* in A minor, the initial 8 bars basically consist of 2 phrases with 4 bars each (Fig. 1B; Supplementary_Material_1 is provided for the actual auditory stimulus). In regard to the first phrase, the melody structure of the last 3 bars is made up of repeated, similar rhythmic patterns of notes: E-B-B, E-A-A, and E-A-G#. The syntactic structures of the 2 phrases, modified by exchanging the fourth (i.e. the last of the first phrase) and sixth (i.e. the second of the second phrase) bars (Fig. 1C; Supplementary_Material_2), are similar to the following 2 sentences with 4 words each: "He (drank, danced, sang), he then (stayed up)" as original sentences, and "He (drank, danced) *then*; he (*sang*, stayed up)" as modified sentences created by exchanging the italicized fourth and sixth words. Note that the structures denoted by parentheses are altered by the swap.

These unnatural stimuli were played by the pianists without an obvious break or articulation error. Any advanced musician or composer with sufficient knowledge of classical music would be expected to detect these structural changes by ear. However, intermediate music students, especially those who had not taken

the 1-wk training in our experiment, might miss most of the unnatural stimuli. Because these detection processes required deeper contextual comprehension of the polyphonic stimuli, we named the experimental condition the "Context" condition.

By using the Wavelab 10 software (Steinberg Media Technologies GmbH, Hamburg, Germany), we digitized the stimuli (16 bit, 48 kHz, stereo), where the loudness of each recorded piece was equally set to −23 LUFS (loudness units relative to full scale). For the stimuli played by each pianist, we extracted 4 natural stimuli and 6 unnatural stimuli of 18 s each for each of the 4 pieces. For each unnatural stimulus, there was always a corresponding natural stimulus in the same range of a piece. With some overlaps between the stimuli, the whole stimuli set covered all portions of the original pieces, except the recapitulation portion (for II-2 alone). Throughout the experiment, the same natural stimulus appeared less than 3 times, whereas the same unnatural stimulus never appeared twice; this was unknown to the participants. The presentation orders of the stimuli were completely randomized across the 4 pieces. If the excerpt had a break point at its beginning, we added a 2-s fade-in; the end was always a break point, and we added a 2-s fade-out. The onset of each error occurred at 7.3 ± 3.5 s [mean ± SD], which was normally distributed.

As a control condition for the Context condition, we used a "Direction" condition requiring sound localization, also with the error-detection task. From the original non-swapped excerpts, we averaged stereo channels and generated monaural stimuli, which were then presented with a 12-dB decrease in either the left or right side of the channels. This decrease caused the excerpts to be heard from a slightly right- or left-oriented sound source. We separately prepared "unnatural" stimuli, where we switched the sides of the decrease for stereo channels. The onset of each error occurred at 9.0 ± 3.5 s, which was normally distributed.

Error detection on these excerpts under the Direction condition thus required correct judgment of sound localization and controlled as a reference for the basic auditory processes for the musical pieces, as well as decision-making associated with error detection. On the other hand, this condition did not require any familiarization with the pieces themselves. By comparing the Context condition with the Direction condition, we were able to clarify brain activations reflecting training effects. In our previous fMRI study (Suzuki and Sakai 2003), we clarified that the brain activations selective to syntactic judgments were observed equally for normal and anomalous (i.e. grammatical and ungrammatical) sentences. Therefore, we combined trials with both normal and anomalous stimuli and focused on musical judgments themselves by utilizing these conditions, rather than comparing the responses to errors with those to normal excerpts.

During the MR scans, the participants wore an MRI-compatible headphone, VisuaStim Digital (Resonance Technology Inc.), a pair of earmuffs (3 M Peltor), and a pair of earplugs (Earasers, Persona Medical) to reduce the high-frequency noises (>1 kHz) of the scanner. Each participant selected a pair of earplugs sized XL, L, M, or S. The stimulus presentation was controlled by the Presentation software package (Neurobehavioral Systems), which also collected the behavioral data (accuracy and response times [RTs]). Before scanning, we appropriately adjusted the sound level for each participant by presenting the first 18 s of piece I-1.

Task

Under the Context and Direction conditions, we used an error-detection task to require participants to detect unnatural stimuli described above as errors. In each trial under both conditions, participants listened to an excerpt for 18 s, and judged whether there was an unnatural portion in the excerpt. A brief beep (0.2 s, at the pitch of A4/a^1) followed the excerpt, and participants pushed either of 2 buttons (one for natural and the other for unnatural) within 2 s including the beep; this time limit was in accordance with our previous study with an error-detection task with music stimuli (Sakai et al. 2022).

One MR scanning run consisted of 8 trials under the Context condition, as well as 4 trials under the Direction condition (in the order of Direction–Direction–Context–Context–. . .–Context–Direction–Direction). At the start of each Direction trial, we presented a beep (0.5 s, at the pitch of G5/g^2), while at the start of each Context trial, we presented a brief beep (0.2 s, at the pitch of G5/g^2) twice, with an interval of 0.1 s. In each run, we presented 4 natural and 4 unnatural stimuli under the Context condition, as well as 2 natural and 2 unnatural stimuli under the Direction condition. Following 4 runs, the participants took a rest for about 10 min before proceeding to 4 more runs in a day; we obtained structural MRI data right after the final run. Due to a health problem, 1 participant quit the experiments after finishing the first 6 runs; structural data were acquired from this participant, so those data were included in the analyses.

MRI data acquisition and analyses

The following methods conformed to the procedures published previously by our team (Ohta et al. 2017; Tanaka et al. 2017; Tanaka et al. 2019). For the MRI data acquisition, the participant was in a supine position, and her/his head was immobilized inside the radio-frequency coil. The MR scans were conducted on a 3.0 T system, GE Signa HDxt 3.0 T (GE Healthcare, Milwaukee, WI). We scanned 30 axial slices, each 3-mm thick and having a 0.5-mm gap, covering the volume range of −38.5 to +66 mm from the anterior to posterior commissure (AC-PC) line in the vertical direction, using a gradient-echo echo-planar imaging (EPI) sequence [repetition time (TR) = 2 s, echo time (TE) = 30 ms, flip angle (FA) = 78°, field of view (FOV) = 192 × 192 mm^2, resolution = 3 × 3 mm^2]. In a single run, we obtained 123 volumes following 4 dummy images, which allowed for the rise of the MR signals. After completion of the fMRI session, high-resolution T1-weighted images of the whole brain (136 axial slices, 1.0 × 1.0 × 1.0 mm^3) were acquired with a 3-dimensional fast spoiled gradient recalled acquisition in the steady state sequence (TR = 8.5 ms, TE = 2.6 ms, FA = 25°, FOV = 256 × 256 mm^2). These structural images were used for normalizing fMRI data.

The fMRI data were analyzed in a standard manner using SPM12 statistical parametric mapping software (Wellcome Trust Center for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm/) (Friston et al. 1995) implemented on MATLAB (Math Works, Natick, MA). The acquisition timing of each slice was corrected using the middle slice (the 15th slice chronologically) as a reference for the EPI data. We realigned the time-series data in multiple runs to the first volume in all runs. The realigned data were resliced every 3 mm using seventh-degree B-spline interpolation so that each voxel of each functional image matched that of the first volume. We removed a run from 1 participant, which included data with a translation of >2 mm in any of the 3 directions or with a rotation of >1.4° around any of the 3 axes.

After alignment to the AC-PC line, each participant's T1-weighted structural image was coregistered to the mean functional image generated during realignment. The coregistered structural image was spatially normalized to the standard brain space as defined by the Montreal Neurological Institute (MNI), using the "unified segmentation" algorithm with light regularization, which is a generative model that combines tissue

segmentation, bias correction, and spatial normalization in the inversion of a single unified model (Ashburner and Friston 2005). After spatial normalization, the resultant deformation field was applied to the realigned functional imaging data. All normalized functional images were then smoothed by using an isotropic Gaussian kernel of 9-mm full-width at half maximum. Low-frequency noise was removed by high-pass filtering at 1/128 Hz.

In a first-level analysis (i.e. the fixed-effects analysis), each participant's hemodynamic responses induced by the Context condition (separated for the Listen and Read training conditions), as well as the Direction condition, in each run were modeled with a boxcar function with a duration of 16 s, excluding 1 s each from both ends of an excerpt. As a control event without any music stimuli, we separately modeled the 2 s after an excerpt, including a brief beep and a button press. The boxcar function was then convolved with a hemodynamic response function. To minimize the effects of head movement, the 6 realignment parameters obtained from preprocessing were included as a nuisance factor in a general linear model. The images under each of the Context (Listen), Context (Read), and Direction conditions, as well as those for the control event, were then generated in the general linear model for each participant and used for the intersubject comparison in a second-level analysis (i.e. the random-effects analysis) with a flexible factorial option. Other nuisance factors were age, LQ, and gender. To examine activated regions in an unbiased manner, we adopted whole-brain analyses.

A repeated-measures analysis of variance (rANOVA) with t-tests was performed with 2 factors (participant groups × events), the results of which were thresholded at uncorrected $P < 0.001$ for the voxel level and at $P < 0.05$ for the cluster level with the false discovery rate (FDR) correction across the whole brain. Following the typical settings in a flexible factorial design for analyzing brain activations (https://www.researchgate.net/publication/267779738), we assumed an equal variance among participants (independent factor) and an unequal variance among participant groups (independent factor). Since the factor of the events (dependent factor) included the control event, we also assumed an unequal variance among those events. For each contrast we tested, an exclusive mask of each negative activation was applied (uncorrected $P < 0.0001$ for the voxel level). Regarding the anatomical identification of activated regions, we basically used the Anatomical Automatic Labeling (AAL) method (http://www.gin.cnrs.fr/en/tools/aal/) (Tzourio-Mazoyer et al. 2002).

In addition to the whole-brain analyses described above, we used the MarsBaR toolbox (http://marsbar.sourceforge.net/) for each participant and obtained the mean percent signal changes from the local maximum of the left or right F3t/F3O activations. We adopted analyses of a region of interest (ROI) to extract the anterior portion of the right superior temporal gyrus (R. aSTG) from a single cluster extending to frontal regions, by using the AAL mask of "Temporal_Sup_R." For analyses of signal changes, as well as for analyses of behavioral data, we used the R software (https://www.r-project.org/).

Results

The major interest of this experiment was to address which of the Listen and Read training conditions was more effective in judging the congruity of structures and context in music, thereby clarifying any changes caused by those training effects (see Fig. 1), while the same stimuli and tasks were used inside the MR scanner across all the conditions. Under the Context and Direction conditions, we tested error or incongruity detection where participants listened to an excerpt and judged whether there was an unnatural portion in the excerpt. Here we report the differences in behavioral data, and then brain activations, under the training conditions based on different modalities.

Behavioral data reflecting training effects and group differences

To examine differences between the Multi and Mono groups, as well as those under the Listen and Read training conditions, we first compared the accuracy rates and RTs for the error-detection task (Fig. 2, Table 2). For the accuracy rates under the Context condition, we performed a 2-way rANOVA (group [Multi, Mono] × training condition [Listen, Read]), which showed a significant main effect of the training condition ($F[1, 36] = 21.3$, $P < 0.0001$), without either a main effect of group ($F[1, 36] = 3.0$, $P = 0.09$) or interaction of the main effects ($F[1, 36] = 0.07$, $P = 0.8$) (Fig. 2A). We confirmed that both groups showed significantly higher accuracy rates under Listen than under Read (Multi: $t[18] = 4.1$, $P = 0.0007$; Mono: $t[18] = 2.7$, $P = 0.01$). These results are remarkable in that the accuracy rates were enhanced under the Listen condition, reflecting pure differences in training modality.

Moreover, the accuracy rates for the Ref group *without* short-term training were about 60%, which were significantly lower than those under Read for the Mono group ($t[21] = 2.3$, $P = 0.03$; see Fig. 2A). Therefore, the accuracy rates from both groups, which were above 70%, should be regarded as the resultant changes caused by those training effects under Listen and Read.

In contrast, the accuracy rates under the Direction condition, without requiring any familiarization with the pieces, were about 80% for all 3 groups. The accuracy rates for the combined Multi and Mono groups were comparable to those for the Ref group ($t[40] = 0.8$, $P = 0.5$).

Some accuracy rates almost reached ceiling effects, and we thus obtained d' from a Z value of the hit rate (correct detection of *unnatural* stimuli) minus that of the false-alarm rate (incorrect responses to *natural* stimuli). For all of the Multi, Mono, and Ref groups, the resultant d' for each condition closely replicated the accuracy data (Fig. 2B), and the d' values for the Ref group were below 1 (i.e. comparable to the chance level) under the Context condition.

Regarding the RTs for the Multi and Mono groups, neither main effect nor interaction was significant under the Context condition ($P > 0.1$) (Fig. 2C); there was a tendency for the RTs of Multi to be shorter than those of Mono ($t[36] = 1.7$, $P = 0.1$). Next, to clarify group differences with respect to the combination of accuracy rates and RTs, we used the accuracy/RTs ratios (Fig. 2D), which were regarded as a normal distribution in either group or training condition (Shapiro–Wilk tests, $P > 0.5$). Under the Context condition, we observed both significant main effects of *group* ($F[1, 36] = 4.5$, $P = 0.04$) and *training condition* ($F[1, 36] = 17$, $P = 0.0002$), without their interaction ($F[1, 36] = 0.03$, $P = 0.9$). For each group, accuracy/RTs were significantly higher under Listen (Multi: $t[18] = 3.3$, $P = 0.004$; Mono: $t[18] = 2.6$, $P = 0.02$), replicating the results of the accuracy rates. The higher values for Multi indicate the superior performance of this group under the Context condition.

To assess differences in piano-training methods, we compared a group of Suzuki students ($n = 23$ excluding the "Ref" group; 11 in Multi, 12 in Mono) with the other participants ($n = 15$; 8 in Multi, 7 in Mono), as shown in Fig. 2E. The latter group ("Others") had been trained with different methods other than the Suzuki Method, including self-training. With respect to accuracy/RTs, a 3-way rANOVA (group$_1$ [Multi, Mono] × group$_2$ [Suzuki,

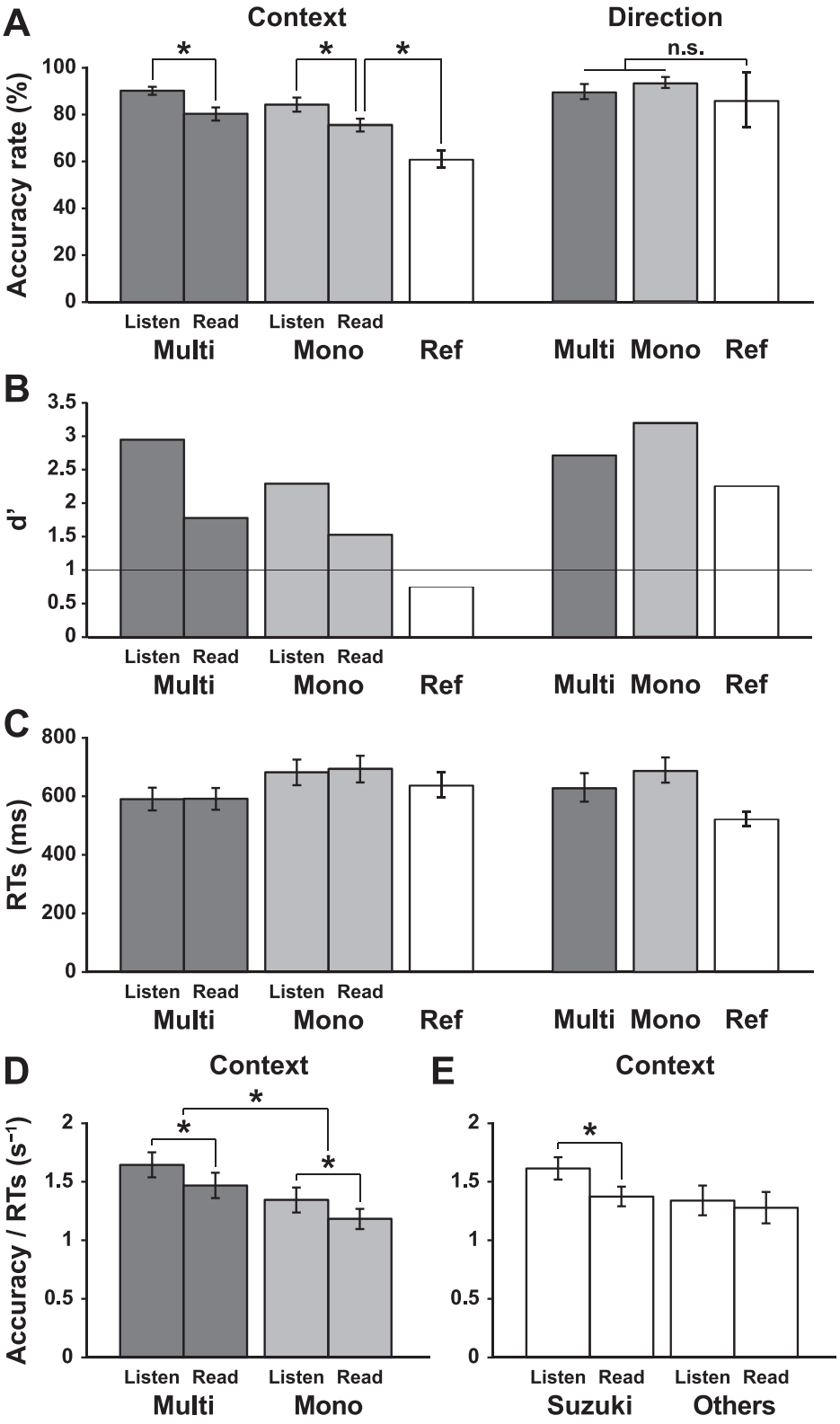

Fig. 2. Behavioral data reflecting training effects and group differences. (A) Accuracy rate in the error-detection task. Under the Context condition, accuracy rates under the Listen training condition were significantly higher than those under the Read training condition, reflecting pure differences in training modality. These differences were observed for both the Multi-instrumentalist ("Multi") and Mono-instrumentalist ("Mono") groups. Note that the stimuli and tasks during the MR scans were the same across all the conditions. Moreover, accuracy rates for the reference ("Ref") group *without* short-term training were significantly lower than those for the Mono group under Read, confirming the presence of training effects under the Context condition. Under the "Direction" condition, i.e. a control condition that did not require any training, accuracy rates for the combined Multi and Mono groups were comparable to those for the Ref group. (B) The *d'*-values, which indicated a more robust estimation of performances without ceiling effects. For the Ref group under the Context condition, values were below 1 (i.e. comparable to the chance level). (C) RTs, which were comparable among groups under the Context condition. (D) Ratios of the accuracy to the RTs (accuracy/RTs) under the Context condition, where higher values indicate better task performances. Values for Multi were significantly higher than those for Mono. (E) The accuracy/RTs for the groups of the Suzuki students (see *Introduction*) and the other participants ("Others"). Values were significantly higher under Listen than under Read for the Suzuki group alone. Error bars indicate the SEM. *$P < 0.05$; n.s., not significant ($P > 0.5$).

Table 2. Behavioral data for each group and condition.

	Context					Direction		
	Multi		Mono		Ref	Multi	Mono	Ref
	Listen	Read	Listen	Read				
Accuracy rates (%)	90 ± 1.7	80 ± 2.8	84 ± 3.0	76 ± 2.7	61 ± 4.0	89 ± 3.2	93 ± 2.3	86 ± 12
RTs (ms)	590 ± 39	591 ± 37	682 ± 44	693 ± 46	638 ± 45	627 ± 49	687 ± 43	522 ± 26

Data are shown as the mean ± SEM (see Fig. 2A and C). Ref, reference group (see *Participants*); RTs, response times.

Others] × training condition [Listen, Read]) showed significant main effects of $group_1$ ($F[1, 34] = 5.3$, $P = 0.03$) and training condition ($F[1, 34] = 13.7$, $P = 0.0007$), as well as an interaction of $group_2$ × training condition ($F[1, 34] = 5.1$, $P = 0.03$), while the other main effect of $group_2$ or other interactions were not significant ($P > 0.1$). Accuracy/RTs for the Suzuki group were significantly higher under Listen than under Read ($t[22] = 4.7$, $P = 0.0001$), whereas those for the Others group were comparable ($t[14] = 1.0$, $P = 0.3$). This significant interaction indicated that *short-term* effects under Listen were facilitated qualitatively by *long-term* training effects for the Suzuki students. Moreover, the Others group did not show significantly better performances under Listen than under Read, suggesting that the Listen condition did not provide an obvious advantage due to the use of the same modality (i.e. auditory stimuli) in the task.

We also analyzed whether total practice hours (for both piano and other instruments) explained group differences, examining quantitative and qualitative effects. The median of practicing hours was 3,200 h, and 2 participants were in this median range. Excluding those participants, we compared 2 groups of participants (18 participants each), i.e. with practicing hours *above* or *below* the median. Regarding accuracy/RTs, a 3-way rANOVA ($group_1$ [Multi, Mono] × $group_3$ [Above, Below] × training condition [Listen, Read]) showed significant main effects of $group_1$ ($F[1, 32] = 5.1$, $P = 0.03$) and training condition ($F[1, 30] = 17.2$, $P = 0.0002$), while the other main effect of $group_3$ or any of the interactions were not significant ($P > 0.2$). Aside from the Suzuki Method, long-term training effects on performances observed across Multi and Mono were not explained quantitatively by the factor of practicing hours alone, suggesting the involvement of qualitative differences.

We evaluated whether memory factors or familiarity effects on the stimuli themselves influenced the task performances under Listen or not. We compared the stimuli played by each pianist, where half of the stimuli was familiar (i.e. used for the 5-d training with a CD) and the other half was unfamiliar regarding playing styles. Combining Multi and Mono (i.e. all participants), both accuracy rates and RTs under Listen were comparable between those stimuli (accuracy rates: $t[37] = 0.6$, $P = 0.5$; RTs: $t[37] = 0.5$, $P = 0.7$).

Differences in brain activation patterns among groups and training conditions

Considering the significant behavioral differences between groups, as well as those between training conditions, we next examined the brain activations separately for individual groups and training conditions (Fig. 3, Table 3). In the [Context (Listen) – Direction] contrast for the Multi group, significant activations in the frontal cortex were localized in the L. LPMC and F3op/F3t, showing left-lateralization (Fig. 3A). Additional activations were observed in the medial presupplementary motor area (pre-SMA), anterior cingulate cortex (ACC), thalamus, and midbrain. Moreover, we found activations through the auditory pathway, starting from the inferior colliculus (IC) and medial geniculate nucleus (MGN) and continuing to the right Heschl's gyrus (R. HG).

There were no significant activations for Mono using the same contrast and threshold, reflecting comparable activations for both conditions; a direct comparison between the groups was not significant, either. With a lower threshold (uncorrected $P < 0.01$ for the voxel level, FDR corrected $P < 0.05$ for the cluster level), significant activations for Mono were detected in regions including the *bilateral* LPMC, F3op/F3t, and F3t/F3O, together with the calcarine, lingual gyri, MGN/thalamus, and IC/midbrain. These results of the left-lateralized activations for Multi are in good agreement with the superior performances for Multi (see Fig. 2D).

In the [Context (Read) – Direction] contrast for Multi, we observed significant activations in various regions including the bilateral LPMC, F3op/F3t, F3t/F3O, and aSTG, together with the R. HG, R. STG/MTG, pre-SMA/ACC, caudate, thalamus including the MGN, midbrain including the IC, cerebellum, and cerebellar nuclei (Fig. 3B). The bilateral activations were replicated for Mono, while activations in the R. HG, caudate, thalamus, midbrain, cerebellum, and cerebellar nuclei, including the above-mentioned auditory pathway, were observed only for Multi. Comparing the 2 groups, the R. LPMC activations for Multi were broader and extended to the more dorsal region, whereas the R. F3t/F3O activations for Mono were broader and extended to the more lateral region. For both groups, activations in the bilateral F3t/F3O and aSTG, as well as a number of right temporo-frontal regions, were observed only under the Read training condition.

We further performed a direct comparison between the main conditions, i.e. the [Context (Read) – Context (Listen)] contrast. For both the Multi and Mono groups, significant activations were observed only in the right temporo-frontal regions: R. LPMC, R. F3op/F3t, R. F3t/F3O, and R. aSTG (Fig. 3C). The R. LPMC activations were broader for Multi, and the R. F3t/F3O activations were broader for Mono; the R. STG/MTG activations were selective to Mono. These activations reflect the structural loads in the task additionally required by the Read training condition (see Fig. 2D).

We also assessed activations for the Suzuki and Others groups in the same direct comparison. For the Suzuki group, language areas consisting of the L. LPMC, L. F3op/F3t, L. F3t/F3O, L. aSTG, and L. STG/MTG were all significantly activated, in addition to their right homologs (Fig. 3D). These regions were basically included in the previous contrasts, but the L. STG/MTG was a unique region, which is the *left homolog* of the R. STG/MTG observed for both Multi and Mono. For the Others group, significant activations were not observed, consistent with their comparable performances between Listen and Read (see Fig. 2E). These enhanced activations in language areas for the Suzuki group also reflect the structural loads required by the Read training condition (see Fig. 2E).

To assess individual differences among the participants, we focused on the R. aSTG, the most prominent among the activated regions under Read (see Fig. 3C and *Z* values in Table 3), as well as on the peaks of the bilateral F3t/F3O, the

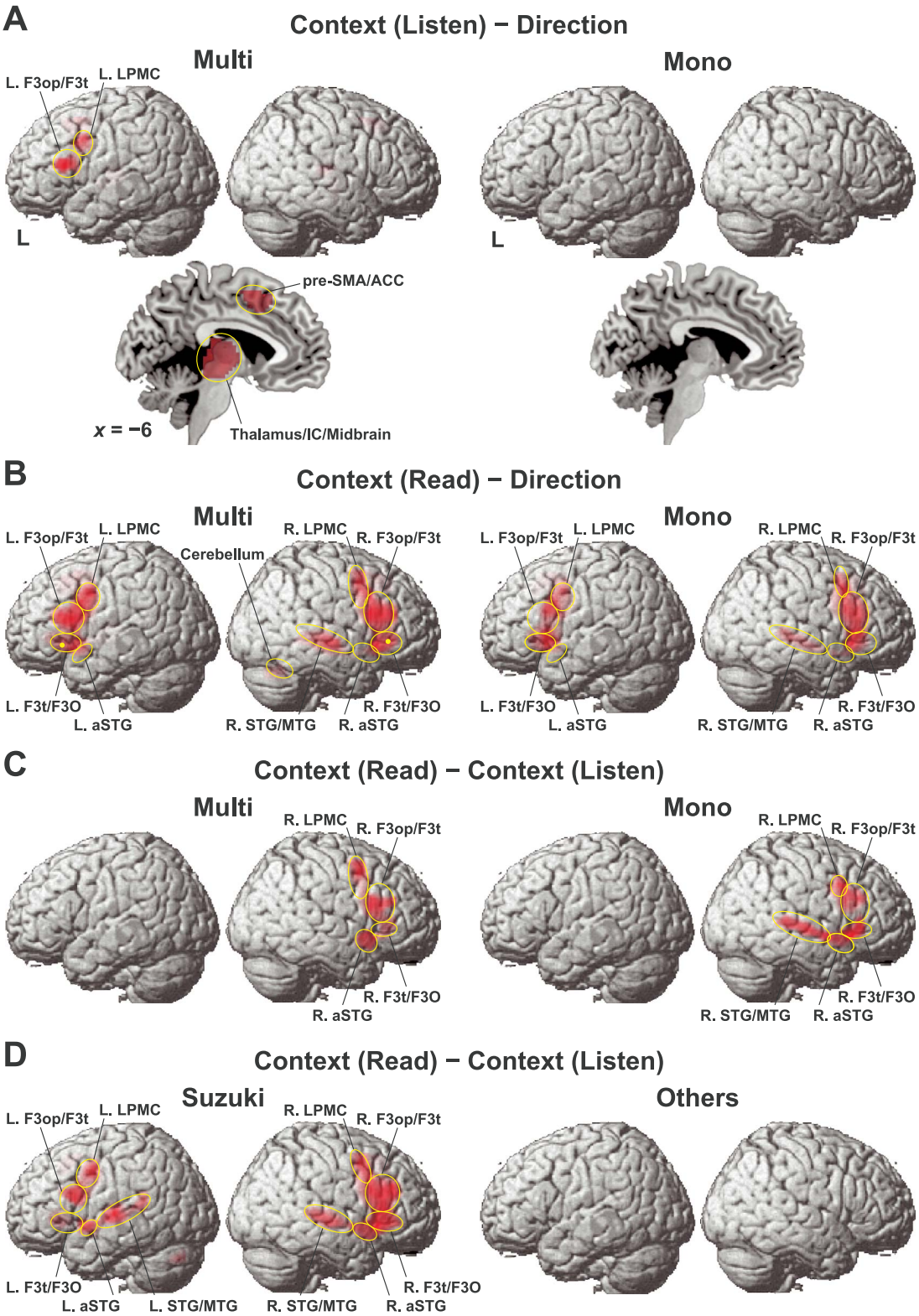

Fig. 3. Differences in brain activation patterns among groups and training conditions. (A) Activations selective to the Context (Listen) condition for the Multi and Mono groups. Left frontal and medial activations were observed specifically for the Multi group, showing a clear left-lateralization and localization of activations. (B) Activations selective to the Context (Read) condition. Similar bilateral activations were observed for both groups. The dots for the Multi group indicate the local maxima of activations in the bilateral triangular/orbital parts of the inferior frontal gyri (F3t/F3O; see Table 3), which were used in correlation analyses (see Fig. 4C). (C) A direct comparison of Context (Read) and Context (Listen) conditions for Multi and Mono. Significant activations were identified in the right frontal and temporal regions, suggesting their supportive roles. Note that activations in the right superior temporal gyrus (R. STG) were localized to the anterior portion (aSTG) for the Multi group, which were used in correlation analyses (see Fig. 4A and B). (D) A direct comparison of Context (Read) and Context (Listen) conditions for the Suzuki and Others groups. Significant activations were observed for the Suzuki group alone in the bilateral frontal and temporal regions. Significance was assigned at FDR corrected $P < 0.05$ for the cluster level.

Table 3. Regions with activations selective to the context comprehension for each group.

Brain Region	BA	Side	Multi					Mono				
			x	y	z	Z	Voxel	x	y	z	Z	Voxel
Context (Listen) – Direction												
LPMC	6/8	L	−48	5	38	3.9	204					
F3op/F3t	44/45	L	−45	20	17	4.0	*					
pre-SMA/ACC	6/32	M	−6	8	56	5.0	228					
			−12	14	38	3.9	*					
Thalamus		M	−12	−22	14	4.2	880					
			12	−19	11	4.1	*					
HG	41	R	33	−28	11	3.8	*					
MGN		M	6	−16	−4	4.5	*					
IC/Midbrain		M	−12	−22	−10	3.9	*					
Context (Read) – Direction												
LPMC	6/8	L	−48	5	35	5.6	1,032	−51	5	44	4.4	884
								−36	−7	26	4.1	*
F3op/F3t	44/45	L	−45	20	17	5.0	*	−39	14	23	4.7	*
F3t/F3O	45/47	L	−30	26	−4	5.4	*	−33	26	−1	4.7	*
aSTG	22	L	−51	8	−7	4.1	*	−51	11	−1	4.1	*
LPMC	6/8	R	48	5	50	4.1	1,556	48	14	41	4.6	1,170
			36	−1	59	3.9	*					
F3op/F3t	44/45	R	39	11	23	6.0	*	45	20	23	5.3	*
F3t	45	R	45	29	23	4.7	*					
F3t/F3O	45/47	R	36	29	−1	4.8	*	33	26	−1	5.1	*
								54	26	8	4.7	*
HG	41	R	33	−28	11	3.8	*					
aSTG	22	R	51	2	−13	3.4	*	51	5	−13	3.6	*
STG/MTG	22/21	R	45	−22	−7	5.8	*	51	−19	−4	4.4	*
			51	−31	−1	5.2	*	54	−37	5	4.0	*
pre-SMA/ACC	6/32	M	−3	8	56	6.4	467	−3	8	53	5.3	332
			−9	14	44	5.2	*	9	17	47	5.3	*
Caudate		L	−15	−4	17	3.4	1,206					
		R	15	8	5	3.7	*					
MGN/Thalamus		M	12	−16	8	4.7	*					
			−9	−13	2	4.6	*					
IC/Midbrain		M	9	−25	−7	4.8	*					
			−9	−25	−10	4.0	*					
Cerebellum VI/Crus I		R	21	−70	−28	3.9	228					
Cerebellum Crus I		R	36	−61	−31	4.1	*					
Cerebellum VIII		M	−3	−64	−37	3.8	*					
Cerebellar nuclei		M	−3	−55	−22	4.0	*					
Context (Read) – Context (Listen)												
LPMC	6/8	R	36	2	56	4.3	731					
		R	51	11	44	3.5	*	54	11	38	3.9	732
F3op/F3t	44/45	R	42	11	23	4.4	*	36	5	32	3.7	*
F3t	45	R	45	32	23	3.6	*	51	26	29	4.0	*
F3t/F3O	45/47	R	39	29	5	4.1	*	39	29	2	3.2	*
								54	26	5	4.0	*
aSTG	22	R	54	8	−7	4.7	*	51	−1	−7	4.4	*
STG/MTG	22/21	R						63	−19	−1	4.3	*
								60	−34	8	4.3	*

Stereotactic coordinates (x, y, z) in the MNI space are shown for activation peaks of Z values, which were more than 16 mm apart (see Fig. 3). A region marked with an asterisk is included within the same cluster as the region in the row right above it. BA: Brodmann's area; L: left; R, right. M, medial. ACC, anterior cingulate cortex. aSTG, anterior portion of the superior temporal gyrus. F3op/F3t/F3O, opercular/triangular/orbital parts of the inferior frontal gyrus. HG, Heschl's gyrus. IC, inferior colliculus. LPMC, lateral premotor cortex. MGN, medial geniculate nucleus. pre-SMA, presupplementary motor area. STG/MTG, superior/middle temporal gyri.

frontal region observed under Read alone (see Fig. 3B and Table 3).

Brain activations related to task performances and group differences

In regard to the R. aSTG activations, which were identified by the [Context (Read) – Context (Listen)] contrast for the Multi group (see Fig. 3C), we were able to adapt ROI analyses, thereby separating activated regions with AAL (see *MRI Data Acquisition and Analyses*). Using the signal changes in the [Context (Read) – Direction] contrast, we observed a significantly *positive* correlation between R. aSTG activations and RTs under the Context (Read) condition for all participants (across both groups) ($r = 0.35$, $P = 0.03$; Fig. 4A), as well as a marginally *negative* correlation between activations and accuracy/RTs ($r = -0.30$, $P = 0.07$; Fig. 4B), while accuracy rates showed no significant correlation ($P = 0.3$). These results indicate that R. aSTG activations reflect additional structural loads for the participants with lower task performances, suggesting individual differences in training effects under Read.

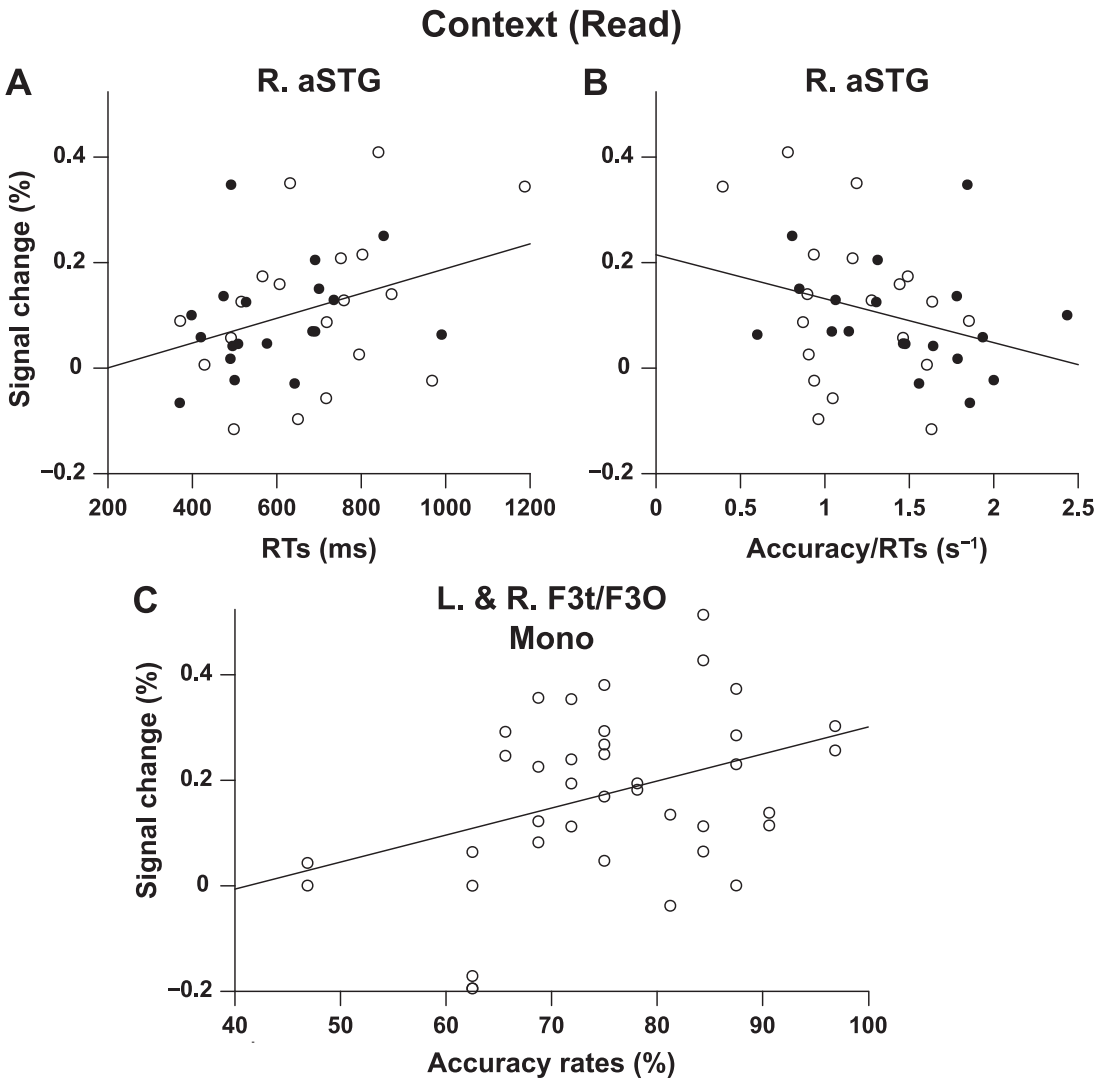

Fig. 4. Brain activations related to task performances and group differences. (A) A positive correlation between the R. aSTG activations and RTs for all participants (Multi: filled dots; Mono: open dots), under the Context (Read) condition. (B) A correlation between the R. aSTG activations and accuracy/RTs for all participants. Note that the performance index of accuracy/RTs makes the correlations reversed. These results indicate that R. aSTG activations reflect additional structural loads for the participants with lower task performances. (C) A positive correlation between the bilateral F3t/F3O activations and accuracy rates for the Mono group alone. The F3t/F3O activations in each hemisphere were treated as independent samples.

Regarding the bilateral F3t/F3O activations, we selected their local maxima based on activations identified by the [Context (Read) – Direction] contrast for Multi (see Fig. 3B): L. F3t/F3O [−30, 26, −4] and R. F3t/F3O [36, 29, −1]. Using the signal changes in each hemisphere as independent samples, the correlations between activation increases and accuracy rates under the Context (Read) condition were not significant for all participants ($r = 0.04$, $P = 0.8$) or for Multi ($r = -0.28$, $P = 0.09$). However, we found a significantly positive correlation for Mono alone ($r = 0.39$, $P = 0.02$) (Fig. 4C). By testing the difference between 2 independent correlations (Zou 2007), we confirmed that the correlation coefficients were significantly different between the 2 groups ($z = 2.9$, $P = 0.004$). These results indicate the different functional roles of the bilateral F3t/F3O and R. aSTG.

Discussion

After a short-term training of either listening to a CD or reading sheet music (Fig. 1A), we compared those training effects by using the error-detection task during MR scans (Fig. 1B and C). The participants were divided into Multi and Mono groups based on their long-term experiences of playing multiple musical instruments vs. a single instrument. We obtained the following results. First, the accuracy rates for error detection became significantly higher under the Listen training condition compared to the Read training condition (Fig. 2A). Moreover, accuracy/RTs for the Multi group were significantly higher than those for the Mono group (Fig. 2D). These behavioral results revealed the clear superiority of *short-term* training by listening to the pieces, and of *long-term* training by playing multiple instruments, when judging structures and context in music. Furthermore, accuracy/RTs were significantly higher under Listen than under Read for the Suzuki students alone (Fig. 2E), consistent with the long-term effects of the Suzuki Method. Our results indicate the presence of long-term training effects, such that playing multiple instruments enhances cumulative effects in music interpretation. Secondly, we observed left-lateralized activations in the L. LPMC and L. F3op/F3t under Listen for the Multi group (Fig. 3A), whereas bilateral activations including the F3t/F3O were observed under Read for both groups (Fig. 3B). Such additional activations were prominent in the right

cortical regions, especially the R. aSTG (Fig. 3C). For the Suzuki students, not only right fronto-temporal activations but *left* activations in the language areas were observed in the direct contrast of [Context (Read) – Context (Listen)] (Fig. 3D). Thirdly, focusing on individual differences among the participants under Read, we found that higher activations in the R. aSTG reflected the structural loads required for participants with lower task performances of all participants (Fig. 4A and B). Moreover, we observed a positive correlation between the signal changes in the bilateral F3t/F3O and the accuracy rates under Read for Mono (Fig. 4C). These additional activations suggest the supportive and differential roles of the bilateral F3t/F3O and right temporal regions. Overall, our results elucidate how the neural substrates of musical judgments reflecting the various effects (e.g. structural processing and understanding in music) are influenced by training.

Our present findings regarding music can be naturally extended to language acquisition. Regarding L1 acquisition, infants without auditory disabilities rely on speech sounds, where motherese (infant-directed speech) plays a major role (Fernald 1985; Werker et al. 1994; Kuhl 2004). The emphasized phrasal or articulatory information in motherese becomes crucial for building syntactic structures among phrases; those structures are the basis for the identification of sounds and meanings/contexts of given speech (Chomsky 1995). On the other hand, phrasal or articulatory information is mostly lost in written words, and thus one has to learn to associate them with speech sounds, syntactic features, and meanings. These learning abilities are regarded as independent of the faculties necessary for L1 acquisition. Similarly, in L2 acquisition at school, the naturalness of sounds tends to be neglected in favor of text comprehension, such that a student learns her/his L2 mostly with written words. Nevertheless, proficiency in L2 listening and reading has been shown to correlate with L1 listening comprehension abilities, indicating a common capacity for L1 and L2 (Vandergrift 2006; Edele and Stanat 2016). Indeed, knowledge of any previously acquired languages can facilitate subsequent language acquisition, a phenomenon theorized by the Cumulative-Enhancement model (CEM) (Flynn et al. 2004; Flynn 2021). In the CEM, linguistic knowledge in multiple languages enhances stepwise cumulative effects in making syntactic structures of sentences, even if used words are completely different among languages. Our previous studies on acquisition in third and fourth languages revealed that complex sentence structures in a novel natural language (Kazakh in our study) were successfully acquired by merely listening to spoken sentences (Umejima et al. 2021, 2024). In general, the superiority of speech sounds over written words is thus clear, and solely text-based learning has a number of problems in language acquisition.

Functional lateralization of language areas was clearly demonstrated by activations in the *left* frontal regions of the L. LPMC and L. F3op/F3t, which were localized under Listen for the Multi group (Fig. 3A). These regions have previously been identified as "grammar centers" for language (Sakai 2005), and they have been shown to be critical for syntactic processing in sentence comprehension (Hashimoto and Sakai 2002; Kinno et al. 2008; Kinno et al. 2014). The grammar centers are actually portions of 3 syntax-related networks (Kinno et al. 2014). Among the left frontal regions activated under Read for both groups (Fig. 3B), the L. F3op/F3t, L. LPMC, and L. F3t/F3O are the key linguistic centers of the networks I, II, and III, respectively; the pre-SMA belongs to the network I. Recently, by increasing syntactic loads, we identified a fourth syntax-related network, as well as additional regions to be included in the original 3 networks (Tanaka et al. 2020). Among the regions selective to Multi in the same contrast, the midbrain, cerebellum, and cerebellar nuclei belong to network II, whereas the thalamus belongs to network IV. This correspondence of activation patterns suggests common neural substrates for processing musical and linguistic structures. Moreover, the IC, MGN, and R. HG, which make up the central auditory pathway (Pickles 2015), were selectively activated for participants of the Multi group under both Listen and Read, indicating the enhancement of lower-order auditory processes, which may subserve the precise identification of tonal features during musical judgments.

The supportive roles of the *right* frontal regions are suggested by activations in the direct contrast of [Context (Read) – Context (Listen)] for both groups (Fig. 3C), because the Read condition was more demanding than the Listen condition, as shown by the behavioral results (Fig. 2). This possibility is in agreement with our previous studies with L2 acquisition, where R. LPMC and R. F3op/F3t/F3O activations were prominent during grammatical processing in L2 but least during grammatical processing in L1 (Sakai et al. 2004; Tatsuno and Sakai 2005). Greater activations of the right frontal regions were also observed during the processing of artificial sentences compared to the processing of natural sentences (Tanaka et al. 2019). In a similar manner, our recent study on multilingualism showed that those right frontal regions were additionally recruited in bilinguals who managed to acquire a third new language at an early stage, but were not recruited in multilinguals with better performances (Umejima et al. 2021). These findings thus specify the necessary conditions for the involvement of the right frontal regions.

The finding of differential correlations between the bilateral F3t/F3O and R. aSTG (Fig. 4) was reminiscent of our previous findings on L2 acquisition—namely, we previously observed that the ventral L. F3t activations for early starters (i.e. participants with longer DOE) were negatively correlated with the accuracy of syntactic decisions on sentences, while these 2 parameters were positively correlated for late starters (Sakai et al. 2009). Moreover, in other reports, we observed a negative correlation between these 2 variables for elder students (Tatsuno and Sakai 2005), and a positive correlation for younger students (Sakai et al. 2004). These findings provided evidence for multiphase changes of activations during L2 acquisition (see *Introduction*). In the present study, we observed that brain activations in the R. aSTG were *saved* for the participants with higher performances from both groups (accuracy/RTs; see Fig. 4B), which would have reflected long-term training effects for several years mostly on auditory processes. This result was consistent with a fall in activations during consolidation of linguistic competence. On the other hand, we found a positive correlation between the bilateral F3t/F3O activations and performances (accuracy; see Fig. 4C) for Mono alone, which would have reflected short-term training effects. Because the Mono group was less experienced regarding the enhancement of cumulative effects in music interpretation than the Multi group, Mono would have been at the initial stage of music acquisition. It is thus possible to extend the multiphase changes of activations found in language acquisition to the case of music acquisition as well, with saved cortical activations similarly reflecting language and musical expertise.

Regarding the group differences we observed, we discuss about 3 issues. The first one is the possibility that participants whose musical decisions were highly accurate tended to play multiple instruments. To exclude this possibility, longitudinal studies examining performance improvements after starting multiple instruments would be necessary, similar to a previous longitudinal study with structural MRI, which found changes in

hippocampal volumes for qualified trainees who aimed to become licensed London taxi drivers (Woollett and Maguire 2011). Interestingly, structural changes in a brain region have also been reported during the training of jugglers (Draganski et al. 2004). It is important to determine how the multiphase changes for cortical activations lead to such anatomical changes. The second issue is whether other factors such as socioeconomic status and the music environment of practicing multiple instruments influenced brain activations. However, according to a questionnaire we conducted, all participants (except 1 in Multi and 2 in Mono) had a habit of both listening to recorded sources and reading sheet music for a newly trained composition. The third issue is how short-term training is transferable to skills in performing or learning music. In the present study, we tested error or incongruity detection in a learned set of stimuli, which would be required for performances as exact reproduction. We acknowledged the distance between music learning in general and our task with an over-learned set of stimuli. To display skilled performances, musicians should acquire competence to interpret the pieces properly, and have enough physical techniques to play as they wish. As regards listening to recorded sources vs. reading sheet music, the ultimate aim in music performance is not to replicate another's interpretation, but is to use the others' recordings and/or the sheet music as a foundation for a musician's own interpretation of the piece. Moreover, a skilled musician will be able to internally "play" the music in aid of practice to realize skillful performances (Lotze et al. 2003; Keller 2012). We did not record the participants' playing of the pieces or collect data on how well they felt learning the pieces, which would partially address these points in future studies.

Lastly, we should consider the distinction between the modalities of music training, i.e. training using sheet music or training through listening by ear. Retrieving vocal or instrumental sounds from sheet music requires the transformation of each musical note to a sound representation with an appropriate pitch and length. Multiple notes are then associated with specified "tempo" or agogic, which plays the basic role of speed control in music, reflecting performing styles and/or emotional states. Individual notes are also linked to "stress" or dynamic accentuation, which is similar to accents in English and other languages, in that those sounds require certain forces to be produced. Correspondingly, the "articulation" of those notes further depends on phrasal structures and their contextual interpretation in music. These processes involved in reading sheet music are thus so demanding and skill-dependent that the performances under Read were lower than those under Listen. We should also note that, while it might be possible to measure brain activations when participants are reading sheet music, skills for making sound representations and mere pictorial memory of the music symbols or layers would provide additional confounding factors. We thus tested auditory stimuli alone in the task and minimized those factors involved; we will leave the testing visual stimuli in the task for future studies. It is striking that the performance differences between Listen and Read were discernible over just 1 wk of training. According to studies on perceptual learning, knowledge can be consolidated and even enhanced by rapid-eye-movement sleep after initial training (Karni et al. 1994). Moreover, the correspondence between the activation patterns observed in participants listening to music (our present study) and participants listening to speech (our above-mentioned previous studies) indicate that phrasal structures in music and language are processed in the brain in a similar manner, as far as inputs are naturally provided. In other words, when it comes to catching phrasal structures, the ear is mightier than the sheet music.

Acknowledgments

We would like to thank the music teachers and participants who have supported our research, especially Sakiko Ishikawa for the use of her recorded performances, as well as Kayono Nagata, Etsuko Suehiro, Yoshihiko Terada, Kanako Nishida, Reiji Inda, Wakana Miyachi, and Mio Noguchi for coordinating the Suzuki-Brain project. We also thank Taichi Nakaza and Naoko Komoro for technical assistance and Hiromi Matsuda for administrative assistance. For the recruitment of participants, we used the website https://www.jikken-baito.com/.

Author contributions

Reiya Horisawa (Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing), Keita Umejima (Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing), Seizo Azuma (Conceptualization, Project administration, Resources, Writing—review & editing), Takeaki Miyamae (Investigation, Writing—review & editing), Ryugo Hayano (Project administration, Writing—review & editing), Kuniyoshi L Sakai (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing).

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Funding

This study received partial funding from the Suzuki School of Music (the Talent Education Research Institute). The funder was not involved in the collection, analysis, interpretation of data, or the decision to submit the study for publication. This study was also supported by Grants-in-Aid for Early-Career Scientists (No. 24K16045) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Conflict of interest statement: We declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability

The authors may make the stimuli and/or experimental paradigm available upon request. The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

Ashburner J, Friston KJ. 2005. Unified segmentation. *NeuroImage*. 26: 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018.

Bernstein L. 1976. *The unanswered question: six talks at Harvard*. Harvard University Press, Cambridge, MA.

Chomsky N. 1995. *The minimalist program*. The MIT Press, Cambridge, MA.

Cross I. 2014. Music and communication in music psychology. *Psychol Music*. 42:809–819. https://doi.org/10.1177/0305735614543968.

Dart T. 1954. *The interpretation of music*. Hutchinson's University Library, London.
Draganski B et al. 2004. Changes in grey matter induced by training. *Nature*. 427:311–312. https://doi.org/10.1038/427311a.
Edele A, Stanat P. 2016. The role of first-language listening comprehension in second-language reading comprehension. *J Educ Psychol*. 108:163–180. https://doi.org/10.1037/edu0000060.
Fernald A. 1985. Four-month-old infants prefer to listen to motherese. *Infant Behav Dev*. 8:181–195. https://doi.org/10.1016/S0163-6383(85)80005-9.
Flynn S. 2021. Microvariation in multilingual situations: the importance of property-by-property acquisition: pros and cons. *Second Lang Res*. 37:481–488. https://doi.org/10.1177/0267658320945761.
Flynn S, Foley C, Vinnitskaya I. 2004. The cumulative-enhancement model for language acquisition: comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses. *Int J Multiling*. 1:3–16. https://doi.org/10.1080/14790710408668175.
Friston KJ et al. 1995. Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp*. 2:189–210. https://doi.org/10.1002/hbm.460020402.
Hashimoto R, Sakai KL. 2002. Specialization in the left prefrontal cortex for sentence comprehension. *Neuron*. 35:589–597. https://doi.org/10.1016/s0896-6273(02)00788-2.
Jackendoff R, Lerdahl F. 2006. The capacity for music: what is it, and what's special about it? *Cognition*. 100:33–72. https://doi.org/10.1016/j.cognition.2005.11.005.
Karni A, Tanne D, Rubenstein BS, Askenasy JJM, Sagi D. 1994. Dependence on REM sleep of overnight improvement of a perceptual skill. *Science*. 265:679–682. https://doi.org/10.1126/science.8036518.
Keller PE. 2012. Mental imagery in music performance: underlying mechanisms and potential benefits. *Ann N Y Acad Sci*. 1252:206–213. https://doi.org/10.1111/j.1749-6632.2011.06439.x.
Kennan K. 1998. *Counterpoint: based on eighteenth-century practice*. Prentice Hall, Upper Saddle River, NJ.
Kinno R, Kawamura M, Shioda S, Sakai KL. 2008. Neural correlates of noncanonical syntactic processing revealed by a picture-sentence matching task. *Hum Brain Mapp*. 29:1015–1027. https://doi.org/10.1002/hbm.20441.
Kinno R, Ohta S, Muragaki Y, Maruyama T, Sakai KL. 2014. Differential reorganization of three syntax-related networks induced by a left frontal glioma. *Brain*. 137:1193–1212. https://doi.org/10.1093/brain/awu013.
Kramer L. 2010. *Interpreting music*. University of California Press, Berkeley, CA.
Kuhl PK. 2004. Early language acquisition: cracking the speech code. *Nat Rev Neurosci*. 5:831–843. https://doi.org/10.1038/nrn1533.
Lerdahl F, Jackendoff R. 1983. *A generative theory of tonal music*. The MIT Press, Cambridge, MA.
Lotze M, Scheler G, Tan HRM, Braun C, Birbaumer N. 2003. The musician's brain: functional imaging of amateurs and professionals during performance and imagery. *NeuroImage*. 20:1817–1829. https://doi.org/10.1016/j.neuroimage.2003.07.018.
Ohta S, Koizumi M, Sakai KL. 2017. Dissociating effects of scrambling and topicalization within the left frontal and temporal language areas: an fMRI study in Kaqchikel Maya. *Front Psychol*. 8:748. https://doi.org/10.3389/fpsyg.2017.00748.
Oldfield RC. 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. 9:97–113. https://doi.org/10.1016/0028-3932(71)90067-4.
Pickles JO. 2015. Auditory pathways: Anatomy and physiology. In: *Handbook of clinical neurology* Aminoff MJ, Boller F, Swaab DF (eds). Amsterdam, Netherlands: Elsevier, pp 3–25.
Sakai KL. 2005. Language acquisition and brain development. *Science*. 310:815–819. https://doi.org/10.1126/science.1113530.
Sakai KL, Miura K, Narafu N, Muraishi M. 2004. Correlated functional changes of the prefrontal cortex in twins induced by classroom education of second language. *Cereb Cortex*. 14:1233–1239. https://doi.org/10.1093/cercor/bhh084.
Sakai KL et al. 2009. Distinct roles of left inferior frontal regions that explain individual differences in second language acquisition. *Hum Brain Mapp*. 30:2440–2452. https://doi.org/10.1002/hbm.20681.
Sakai KL, Oshiba Y, Horisawa R, Miyamae T, Hayano R. 2022. Music-experience-related and musical-error-dependent activations in the brain. *Cereb Cortex*. 32:4229–4242. https://doi.org/10.1093/cercor/bhab478.
Savage PE et al. 2021. Music as a coevolved system for social bonding. *Behav Brain Sci*. 44:e59. https://doi.org/10.1017/S0140525X20000333.
Steinschaden B, Zehetmair H. 1985. *Ear training and violin playing*. Summy-Birchard Inc., Alfred Publishing Co. Inc, Van Nuys, CA.
Suzuki S. 2013. *Nurtured by love (revised edition)*. Alfred Publishing Co. Inc, Van Nuys, CA.
Suzuki K, Sakai KL. 2003. An event-related fMRI study of explicit syntactic processing of normal/anomalous sentences in contrast to implicit syntactic processing. *Cereb Cortex*. 13:517–526. https://doi.org/10.1093/cercor/13.5.517.
Tanaka K, Ohta S, Kinno R, Sakai KL. 2017. Activation changes of the left inferior frontal gyrus for the factors of construction and scrambling in a sentence. *Proc Jpn Acad Ser B Phys Biol Sci*. 93:511–522. https://doi.org/10.2183/pjab.93.031.
Tanaka K et al. 2019. Merge-generability as the key concept of human language: evidence from neuroscience. *Front Psychol*. 10:2673. https://doi.org/10.3389/fpsyg.2019.02673.
Tanaka K, Kinno R, Muragaki Y, Maruyama T, Sakai KL. 2020. Task-induced functional connectivity of the syntax-related networks for patients with a cortical glioma. *Cereb Cortex Commun*. 1:tgaa061. https://doi.org/10.1093/texcom/tgaa061.
Tatsuno Y, Sakai KL. 2005. Language-related activations in the left prefrontal regions are differentially modulated by age, proficiency, and task demands. *J Neurosci*. 25:1637–1644. https://doi.org/10.1523/jneurosci.3978-04.2005.
Tzourio-Mazoyer N et al. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*. 15:273–289. https://doi.org/10.1006/nimg.2001.0978.
Umejima K, Flynn S, Sakai KL. 2021. Enhanced activations in syntax-related regions for multilinguals while acquiring a new language. *Sci Rep*. 11:7296. https://doi.org/10.1038/s41598-021-86710-4.
Umejima K, Flynn S, Sakai KL. 2024. Enhanced activations in the dorsal inferior frontal gyrus specifying the who, when, and what for successful building of sentence structures in a new language. *Sci Rep*. 14:54. https://doi.org/10.1038/s41598-023-50896-6.
Vandergrift L. 2006. Second language listening: listening ability or language proficiency? *Mod Lang J*. 90:6–18. https://doi.org/10.1111/j.1540-4781.2006.00381.x.
Werker JF, Pegg JE, McLeod PJ. 1994. A cross-language investigation of infant preference for infant-directed communication. *Infant Behav Dev*. 17:323–333. https://doi.org/10.1016/0163-6383(94)90012-4.
Woollett K, Maguire EA. 2011. Acquiring "the knowledge" of London's layout drives structural brain changes. *Curr Biol*. 21:2109–2114. https://doi.org/10.1016/j.cub.2011.11.018.
Zou GY. 2007. Toward using confidence intervals to compare correlations. *Psychol Methods*. 12:399–413. https://doi.org/10.1037/1082-989X.12.4.399.

Cerebral Cortex, 2025, **35**, bhaf072
https://doi.org/10.1093/cercor/bhaf072
Original Article