Behavior Research Methods (2021) 53:2007–2024
https://doi.org/10.3758/s13428-020-01528-8

The Musical Ear Test: Norms and correlates from a large sample of Canadian undergraduates

Swathi Swaminathan [1,2] · Haley E. Kragness [3,4] · E. Glenn Schellenberg [3,5]

Accepted: 16 December 2020 / Published online: 11 March 2021

Abstract

We sought to establish norms and correlates for the Musical Ear Test (MET), an objective test of musical ability. A large sample of undergraduates at a Canadian university ($N > 500$) took the 20-min test, which provided a Total score as well as separate scores for its Melody and Rhythm subtests. On each trial, listeners judged whether standard and comparison auditory sequences were the same or different. Norms were derived as percentiles, *Z*-scores, and *T*-scores. The distribution of scores was approximately normal without floor or ceiling effects. There were no gender differences on either subtest or the total score. As expected, scores on both subtests were correlated with performance on a test of immediate recall for nonmusical auditory stimuli (Digit Span Forward). Moreover, as duration of music training increased, so did performance on both subtests, but starting lessons at a younger age was not predictive of better musical abilities. Listeners who spoke a tone language exhibited enhanced performance on the Melody subtest but not on the Rhythm subtest. The MET appears to have adequate psychometric characteristics that make it suitable for researchers who seek to measure musical abilities objectively.

Keywords music · aptitude · expertise · training · melody · rhythm

Over the past couple of decades, it has become well established that musical ability is correlated with many musical and nonmusical abilities (e.g., Bidelman, Hutka, & Moreno, 2013; Piro & Oritz, 2009; Schellenberg, 2006), as well as with functional and structural differences in the brain (for review see Herholz & Zatorre, 2012). Although the roles of nature and nurture in determining musical ability and who takes music lessons remain in doubt (for reviews see Schellenberg, 2020; Swaminathan & Schellenberg, 2019), accurate and objective measurement of musical ability is essential for clarifying these issues. In the present investigation, we administered a measure of musical ability (or musical competence) to a large sample of Canadian undergraduates. Our goals were to develop norms that could be used for interpreting performance levels in future research, and to document other individual differences that vary in tandem with musical ability.

In principle, individual differences in musical ability among participants with no music training must stem from differences in predispositions for musical ability, or musical *aptitude* (natural musical ability), assuming that all other environmental factors are equal. It is not surprising, then, that the history of musical-aptitude testing has been influenced by changes over time in scholars' attitude toward the concepts of aptitude and talent, and toward nativism more generally. In the early 1900s, individual differences in aptitude were assumed to be real and the first modern tests of aptitude were developed. Later in the 20th century, however, the notion of talent was questioned (Ericsson, Rampe, & Tesch-Römer, 1993; Howe, Davidson, & Sloboda, 1998), such that individual differences in musical achievement were considered to be the consequence of practice and other environmental factors such as parental support and encouragement. Although this view was consistent with the dominance of behaviorism in experimental psychology in the early 20th century (Graham, 2019; Skinner, 1976), it lingered long after the birth of

✉ E. Glenn Schellenberg
g.schellenberg@utoronto.ca

[1] Department of Psychology, University of Toronto, Toronto, Canada

[2] Brain and Mind Institute, University of Western Ontario, London, Canada

[3] Department of Psychology, University of Toronto Mississauga, Mississauga, Canada

[4] Department of Psychology, University of Toronto Scarborough, Scarborough, Canada

[5] Centro de Investigação e Intervenção Social (CIS-IUL), Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal

cognitive science and the so-called cognitive revolution (Thagard, 2019). By the early 21st century, however, results from studies of behavioral genetics implicated a role for genes in virtually every measurable human characteristic (DiLalla, 2017), such that there was a resurgence of interest in individual differences in musical aptitude, and in measuring these differences accurately.

As noted, before this time many scholars considered musical talent and aptitude to be something of a myth, or at the very least unimportant theoretically, such that musical accomplishments were attributed primarily to *deliberate practice* (Ericsson, 2006; Ericsson et al., 1993), a view that is now considered to be non-defensible (Hambrick, Macnamara, Campitelli, Ullén, & Mosing, 2016; Ullén, Hambrick, & Mosing, 2016). Indeed, approximately 10,000 hours of practice were considered to be the principal determinant of notable accomplishment in music and other domains, which almost anyone could achieve with the requisite amount of time and effort. This perspective became part of folk wisdom, such that a joke of unknown origin (Question: *How do you get to Carnegie Hall?* Answer: *Practice*) is mentioned on the Carnegie Hall website (https://www.carnegiehall.org/Blog/2016/04/The-Joke).

As it turns out, the association between practice and achievement remains valid, but the link is weaker than once thought, for music as well as for other domains such as sports, games, education, and professions (Macnamara, Hambrick, & Oswald, 2014; Macnamara, Moreau, & Hambrick, 2016; Meinz & Hambrick, 2010). Moreover, in the case of musical achievement, twin studies demonstrate that practice does not influence everyone equally. Rather, practice has a stronger effect for those born with the genetic propensities for musicianship (Hambrick & Tucker-Drob, 2015; Mosing, Madison, Pedersen, & Ullén, 2016). Twin studies further document that genetic factors play a role in determining a musician's choice of musical genre and instrument (Mosing & Ullén, 2018), and, most crucially, amount of *practice* (Butkovic, Ullén, & Mosing, 2015; Mosing, Madison, Pedersen, Kuja-Halkola, & Ullén, 2014). In other words, the *sine qua non* of an environmental contribution to musical accomplishment is actually a gene–environment interaction (Hambrick & Tucker-Drob, 2015).

One of the first tests of musical aptitude was created by Seashore (1919; Seashore, Saetveit, & Lewis 1960). The goal of Seashore's test and others that followed was to determine whether an individual—typically a child—was a suitable candidate for music training. Perhaps the most widely used test in North America was Gordon's *Musical Aptitude Profile* (Gordon, 1965), although tests by Wing (1962) and Bentley (1966) were common in the UK. The various tests all had at least two subtests, one measuring pitch (melody) perception and another measuring temporal (rhythm) perception. The tests of Seashore and Gordon actually had six and seven subtests, respectively, which included measures of other musical dimensions such as the perception of loudness and meter. Gordon subsequently simplified his approach in his *Measures of Music Audiation*, which were available in *Primary* (Grades K-3, Gordon, 1979), *Intermediate* (Gordon, 1982, Grades 1–6), and *Advanced* (Grades 7–Adult, Gordon, 1989) versions. Each of the three versions provided separate scores for melody and rhythm. *Audiation* was a term coined by Gordon (1979) to describe the process of retaining and comparing two musical sequences—a standard followed by a comparison—presented in succession. As such, audiation relies on short-term memory for musical stimuli.

Since 2010, several new measures of musical expertise were developed, including the Musical Ear Test[1] (MET, Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010a), the Profile of Music Perception Skills (PROMS, Law & Zentner, 2012), the Swedish Musical Discrimination Test (SMDT, Ullén, Mosing, Holm, Eriksson, & Madison, 2014), the Harvard Beat Assessment Test (H-BAT, Fujii & Schlaug, 2013), and the Goldsmiths Musical Sophistication Index (Gold-MSI, Müllensiefen, Gingras, Musil, & Stewart, 2014). Because musical ability is now accepted to be a consequence of both genetic and environmental influences, the term *aptitude* has fallen out of use in favor of more neutral terms such as *ability* or *competence*. One notable exception is the Montreal Battery for the Evaluation of Amusia (MBEA, Peretz, Champod, & Hyde, 2003), which has the goal of diagnosing *congenital amusia*—musical abilities that are congenitally and unusually low.

In our laboratory (Swaminathan & Schellenberg, 2017, 2018; Swaminathan, Schellenberg, & Khalil, 2017; Swaminathan, Schellenberg, & Venkatesan, 2018), we opted to use the MET because it is an objective measure of musical ability, freely available, and computer-administered but not self-paced, such that testing time is exactly the same (20 min) for each participant. Like the Gordon (1965, 1979, 1982, 1989) tests, the MET provides separate scores for its Melody and Rhythm subtests, which allowed us to test hypotheses about specific links between rhythm perception and language ability (Swaminathan et al., 2018; Swaminathan & Schellenberg, 2017, 2019), for example, and between melody perception and the ability to speak a tone language (Swaminathan et al., 2018). The MET has additional advantages compared to other recent tests. For example, the original PROMS is much longer in duration (1 h); the SMDT (Melody subtest) requires participants to identify a specific note that is changed on different trials, a task that seems needlessly difficult and molecular for participants with no music training; the H-Bat tests only rhythm perception and uses an adaptive procedure, which means that test duration varies across participants; the Gold-MSI relies on self-reports; and the MBEA is

[1] Freely available from Peter Vuust: petervuust@gmail.com

designed to detect abnormally poor musical abilities, such that most individuals perform at close-to-ceiling levels.

In the present report, we analyzed data from a large sample of participants who took the MET previously in our laboratory as part of a series of studies that examined associations between musical expertise and nonmusical abilities (Swaminathan et al., 2017, 2018; Swaminathan & Schellenberg, 2017, 2018). The large number of participants aggregated across samples allowed us to derive norms, such that raw scores can be interpreted meaningfully in the future, at least in relation to our sample of undergraduates attending a Canadian university. For example, a raw Total score of 81 out of 104 is meaningless on its own, but when converted to a percentile (82nd), *Z*-score (114), or *T*-score (60), this score represents relatively good performance that is approximately one SD above the mean.

We also sought to identify individual-difference variables that predict performance on the MET. The scale's original authors (Wallentin et al., 2010a; Hansen, Wallentin, & Vuust, 2013) reported a positive correlation with Digit Span Forward, a widely used test that measures the capacity of auditory short-term memory (Conway et al., 2005). This association is not surprising because the MET requires same-different judgments on each trial after participants hear a standard and a comparison auditory sequence. We included the entire Digit Span test (Forward and Backward), however, so that we would have an additional, more challenging measure of immediate-memory capacity, one that required participants to re-order the to-be-remembered items. Digit Span Backward is typically considered to measure the capacity of *working* rather than short-term memory (Conway et al., 2005). Based on the structure of the MET and previous findings, we expected that Digit Span Forward would be a particularly good predictor of MET scores. We also included a version of Raven's Advanced Progressive Matrices (Bors & Stokes, 1998; Raven, 1965) as a proxy measure of individual differences in general cognitive but non-verbal (and non-auditory) ability. As with any specific measure of a perceptual or cognitive ability, one would expect musical ability to have a small but reliable correlation with general cognitive ability (Carroll, 1993).

The construct validity of tests that are designed to measure musical ability objectively cannot be assessed directly, because there is no consensus about what musical ability is, what it involves, and how it should be measured. Measures of performance ability or success are inadequate because some individuals have high levels of musical ability but no opportunity to learn how to sing or to play an instrument (Law & Zentner, 2012). Accordingly, available tests tend to focus on perceptual and cognitive abilities. Validity is documented by positive associations with (1) performance on other tests of musical ability (e.g., Law & Zentner, 2012; Peretz et al., 2003) and (2) music training and/or musicianship (Fujii & Schlaug, 2013; Law & Zentner, 2012; Ullén et al., 2014; Wallentin et al., 2010a). The latter association assumes that, on average, performance of a test of musical ability should be better among individuals with music training, whether musical ability promotes the likelihood of music training, music training improves musical abilities, the association is bi-directional, or an unidentified variable is driving the association. In the original report, performance on the MET varied with the participant's status as a musician (professional > amateurs > nonmusician), and it was associated positively with amount of practice (Wallentin et al., 2010a).

In the present sample of Canadian university students, very few were professional musicians. Consequently, we measured duration of music lessons taught privately and in school. In previous studies (Swaminathan & Schellenberg, 2017, 2018), duration-of-training was positively associated with performance on the MET, although this association was stronger for the Melody than for the Rhythm subtest (Swaminathan et al., 2017, 2018). The large sample included here allowed us to test the reliability of these earlier results, and to test specifically whether music training is a better predictor of performance on one of two subtests.

Other goals were to test theories of associations between musical ability and *onset* of training, and between musical ability and language background. Onset of training is important because early childhood may represent a sensitive period during which music training has a more pronounced influence on behavior and brain structure and function (Baer et al., 2015; Bailey & Penhune, 2010, 2012, 2013; Bailey, Zatorre, & Penhune, 2014; Penhune, 2011; Steele, Bailey, Zatorre, & Penhune, 2013; Steele & Zatorre, 2018; Vaquero, Rousseau, Vozian, Klein, & Penhune, 2020; Watanabe, Savion-Lemieux, & Penhune, 2007), in line with principles of plasticity (Penhune, 2019, 2020; Penhune & de Villers-Sidani, 2014). Another possibility, however, is that certain musical predispositions promote musical participation at a young age, as the genetic studies imply (Mosing et al., 2014). Either way, our large sample provided us with a powerful test of whether musical ability tends to be better among individuals who start music training early in life. Nevertheless, if such an association emerged, it would not inform the issue of causal direction.

In previous studies (Swaminathan et al., 2018; Zhang, Xie, Li, Shu, & Zhang, 2020), participants who spoke a tone language outperformed other participants on the Melody subtest but not on the Rhythm subtest of the MET. This finding is consistent with others showing that tone-language use is associated with altered pitch-processing mechanisms in the brain (Bidelman & Chung, 2015; Bidelman & Lee, 2015), and that tone-language speakers have enhanced discrimination, memory, and processing speed for pitch (Bidelman et al., 2013; Hutka, Bidelman, & Moreno, 2015; Stevens, Keller, & Tyler, 2013). More generally, these results are in line with the theoretical proposal that the neural encoding of pitch

works similarly for speech and music, such that experience in one domain (e.g., speaking a tone language, music training) facilitates pitch perception in the other domain (Bidelman, Gandour, & Krishnan, 2011; Wong, Skoe, Russo, Dees, & Kraus, 2007). In the present study, we expected a performance advantage on the Melody subtest among participants who spoke a tone language.

Finally, because the individual-difference variables that we measured would be collinear to some extent, we used multiple regression to determine which ones made independent contributions in predicting musical ability when nonmusical abilities (as measured by Digit Span Backward and Raven's test) were held constant. We expected to find positive partial associations between performance on the MET and music training, between MET scores and scores on Digit Span Forward, and between performance on the Melody subtest and individuals who spoke a tonal language. For onset of music training, however, we were agnostic because onset of training is typically confounded with duration of training. In any event, our large sample afforded ample power to determine which variables were independent predictors of musical ability.

Method

The study protocol was approved by the Research Ethics Board at the University of Toronto.

Participants

The participants were 523 undergraduates (350 women, 168 men, five unreported). Most were registered in an introductory psychology class at a mid-size, suburban campus in Canada and received partial course credit for their participation. The others received token remuneration. The mean age was 19 years (SD = 24 months) but most (n = 391) were between 17 and 19. Participants were originally recruited and tested for studies of music training and nonmusical abilities, including 151 tested by Swaminathan and Schellenberg (2017), 48 from Swaminathan et al. (2017), 165 from Swaminathan et al. (2018), and 53 from Swaminathan and Schellenberg (2018). Others included in the present sample were tested in similar but unpublished studies (n = 67), or excluded from the earlier samples because of failure to meet inclusion criteria for that specific study (e.g., complete data, language background, n = 39).

Most participants had no history of private music lessons (n = 369) but some history of school music lessons (n = 280). On average, they had 2.6 years of private music lessons (SD = 5.4) and 2.5 years of music lessons received in school (SD = 3.8). The average age at which music training began was 10.5 years (SD = 5.5). Most participants were native speakers of English (n = 319), but a sizeable portion (n = 197) had a native language other than English (language data missing for seven participants). Approximately 37% of the nonnative speakers (n = 73) first learned to speak a tone language. Other participants spoke a tone language as a second language or as simultaneous bilinguals, such that the sample comprised 107 tone-language speakers. The broad heterogeneity in language background was commensurate with the multicultural make-up of the local community and the student population.

Measures and procedure

Each participant took the MET individually while sitting in front of a computer in a sound-attenuating booth wearing high-quality headphones. The Melody subtest was administered before the Rhythm subtest. Both subtests had 52 trials (half *same*, half *different*), with sequences presented at 100 beats per minute (bpm). For both subtests, feedback was provided on two initial practice trials but not on the 52 test trials that followed. Participants recorded their responses on an answer sheet with a pen (there was no visual display on the monitor). On each trial, their task was to determine whether the two auditory sequences were the same, and to mark either *YES* or *NO*. Scores were calculated as the number of correct responses.

The same metrical structure (4/4 time) and underlying beat rate, 600 ms per beat, was maintained for both subtests. This tempo, which corresponds to 100 beats per minute (bpm), is slightly slower than the standard dance tempo of 120 bpm (each beat = 500 ms). Throughout both subtests, a metronome sound at the underlying beat level was played at a lower amplitude than the stimuli.

The trial structure was the same for both Melody and Rhythm subtests. For both subtests, the downbeat of each sequence was always 4800 ms (eight beats) after the downbeat of the previous sequence. All Melody sequences were five beats long, and Rhythm sequences were either four, five, or six beats long. A male voice announced the trial number 1200 ms prior to the first downbeat of each trial.

For the Melody subtest, the first and last sounds of each sequence always aligned with the first and last downbeats of the sequence. For the Rhythm subtest, the first sound of a sequence could align with either the first downbeat or a subdivision of the downbeat (e.g., after a half-beat rest). Similarly, the last sound could occur on the last downbeat, a subdivision of that beat, or a subdivision of the penultimate beat. Thus, although the time window between the first downbeat of consecutive sequences was always the same, the amount of time between onset of the *first sound* of a sequence and the *last sound* of the previous sequence on the Rhythm subtest varied slightly (i.e., from 4200–5250 ms between trials and from 1950–2700 ms within a trial).

Sequences in the Melody subtest comprised three to eight grand-piano tones, all of which fell within the range from A3

(3 semitones below middle C) to E5 (16 semitones above middle C). Tones were half-notes (1200 ms), quarter-notes (600 ms), and eighth-notes (300 ms). On all trials, standard and comparison sequences had the same number of tones. On *different* trials (26 of 52), two or three adjacent tones in the standard were reversed in the comparison sequence, or 1–2 tons of the comparison sequence were displaced in pitch relative to the standard, usually by 1–2 semitones (there was one instance of a 5-semitone change). For half of *different* trials, the manipulation changed the contour (i.e., the pattern of upward and downward changes in pitch). In the other half, the contour stayed the same but the intervals (pitch distances) before and after displaced tones were altered. The melodies were in major mode, minor mode, or atonal (i.e., containing non-diatonic tones) on 20, 7, and 25 trials, respectively. Differences among trials in terms of mode and contour contributed to task difficulty and were distributed randomly throughout the Melody subtest. Examples are illustrated with musical notation in Fig. 1. The figure provides two examples of *same* trials, one of which was easy, the other more difficult, and two *different* trials, one easy and one difficult.

On the Rhythm subtest, sequences comprised 4 to 11 wood-block sounds. All wood-block sounds were identical in terms of pitch, but onset-to-onset durations were more variable than in the Melody subtest (i.e., 150, 200, 300, 400, 450, 600, 900, 1200, and 1800 ms). On different trials, the comparison sequence differed from the standard by presenting one or two sounds early or late, presenting one sound early and one sound late, swapping adjacent durations in one or two instances, shifting two to four adjacent sounds ahead or behind in time in an identical manner, or adding an additional sound. When the first or last sound was displaced in time, the comparison sequence had a different onset-to-onset duration between the first and last sounds compared to that of the standard. Examples are provided in Fig. 1, with two *same* trials (one easy, one difficult) and two *different* trials (one easy, one difficult).

The testing context was identical for all participants, who also completed a questionnaire that asked for information about music training and language background. A sizeable majority also provided information about socio-economic status (SES), specifically mother's (n = 440) and father's (n = 442) highest level of education, both measured on 8-point scales (1 = did not finish high school, 8 = graduate degree), and annual family income (n = 431), measured on a 9-point scale (1 = $25,000 or less, 9 = $200,000 or more). A principal component (hereafter, *SES*) was extracted from the three measures for use in the statistical analyses in order to reduce collinearity and measurement-specific error. It accounted for 56.9% of the variance in the original items. Mother's education and father's education loaded highly onto the latent variable (rs > .8); family income had a smaller loading (r = .588). SES scores for participants with missing data were formed by standardizing and averaging the income and education data that were available, such that SES scores were available for 448 participants (86% of the sample).

Questions about language background required participants to identify their native (first) language, and all of the languages they spoke or understood. For each language, they indicated their proficiency on seven 7-point scales (re: speaking, reading, writing, comprehension, vocabulary, fluency, and accent). As noted, most participants completed a test of nonmusical abilities (e.g., speech perception, reading, IQ), the results of which were reported previously (Swaminathan et al., 2017, 2018; Swaminathan & Schellenberg, 2017). A majority of participants (n = 381) also took the Digit Span test and a short version of Raven's Advanced Progressive Matrices (Bors & Stokes, 1998). The Digit Span test provided separate scores for forward and backward span, whereas the Raven's test provided a measure of general but nonverbal cognitive ability. Other participants did not take the Digit Span test, but instead took the complete Raven's test (n = 51) or the short version (n = 24). To equate the short and complete versions of the Raven's test in the statistical analyses, scores were standardized separately for both versions so that they had the same mean and standard deviation (0 and 1, respectively).

Results

Preliminary analysis examined the internal reliability of the MET subtests and Total scores. We calculated Cronbach's alphas and split-half correlations (Spearman–Brown formula) for the whole sample as well as for individuals with no music training and individuals with at least 10 years of training. The results are reported in Table 1. For the whole sample, alphas were lower than those reported by the test developers (Melody: .82, Rhythm: .69, Total: .85; Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010b), which might be expected from a more heterogeneous sample. In general, the entire scale had better internal reliability than the subtests. Alphas were lower for the Rhythm scores than for Melody scores, but split-half correlations were similar. The statistics did not provide evidence that the MET was more reliable for highly trained individuals. Although alphas were higher for these participants than for untrained participants, split-half reliabilities were similar for both groups for the Rhythm subtest, but higher for the untrained group for the Melody subtest and Total scores.

The analyses that follow include standard frequentist statistics, which evaluate the probability of the null hypothesis given the observed data. They also include, whenever possible, Bayesian statistics calculated with JASP 0.10.2 (JASP Team, 2019), using default priors. Bayesian statistics compare the likelihood (or the odds) of the observed data under the alternative compared to the null hypothesis. JASP software

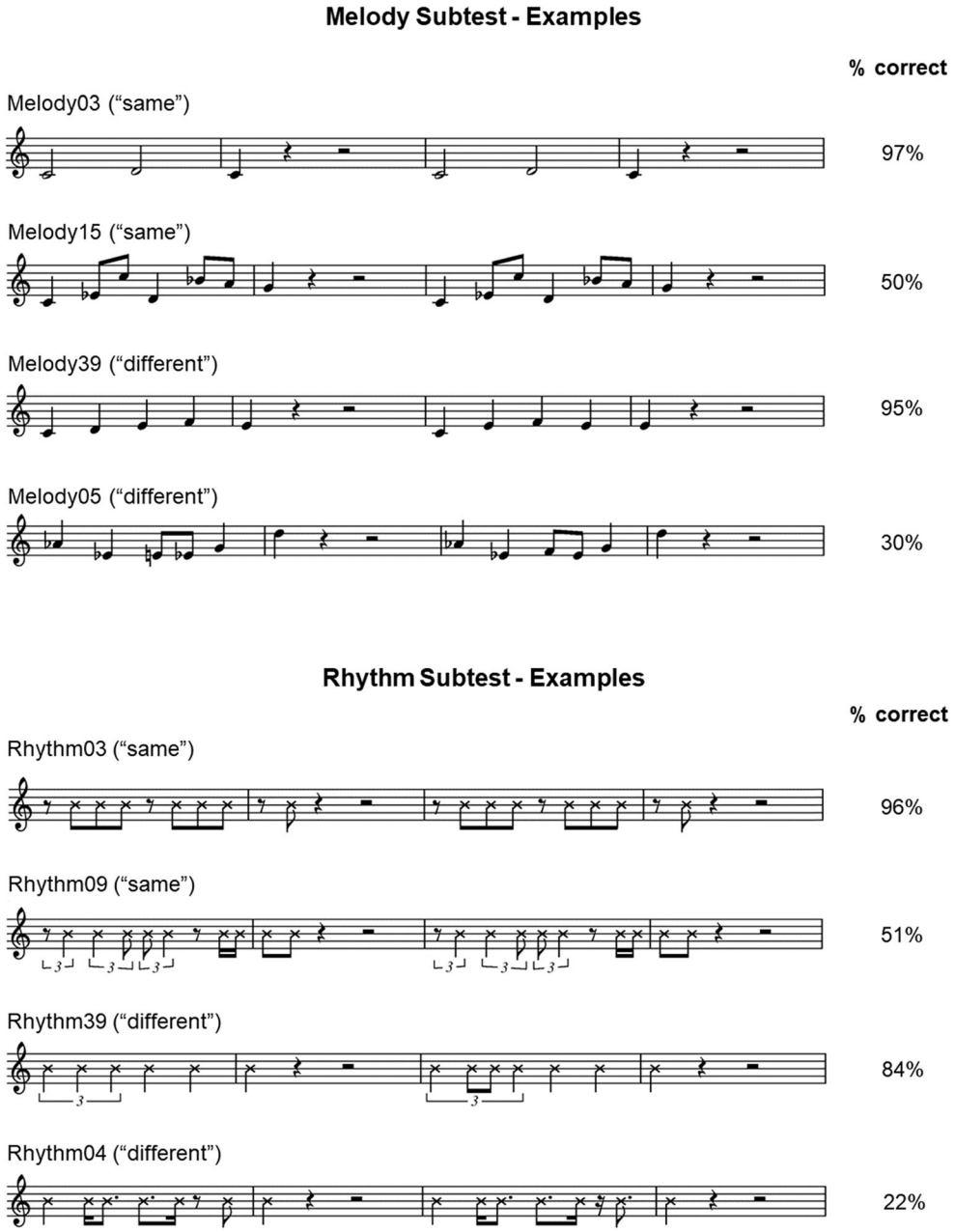

Fig. 1 Examples of trials in musical notation from the Melody (*upper*) and Rhythm (*lower*) subtests of the MET. For both subtests, two same trials and two different trials are illustrated, the first being easy and the second more difficult

includes Bayesian counterparts to paired *t* tests, independent-samples *t* tests, analysis of variance (ANOVA), analysis of covariance (ANCOVA), repeated-measures and mixed-design ANOVAs, as well as correlation and regression, including multiple regression. All tests provide Bayes factors (i.e., BF_{10}, reported here with three-digit accuracy). When a Bayes factor is equal to 1, the observed data are equally likely under the alternative and null hypotheses. When $BF_{10} > 1$, the observed data are more likely under the alternative hypothesis. When $BF_{10} < 1$, the observed data are more likely under the null hypothesis. According to conventional rules of thumb (Jarosz & Wiley, 2014; Jeffreys, 1961), weak or anecdotal evidence for the alternative hypothesis is provided by values between 1 and 3, with values of 3–10, 10–30, 30–100, and over 100 providing substantial, strong, very strong, and decisive evidence, respectively. Conversely, values of 1.0–.33, .33–.10, .10–.03, .03–.01, and less than .01 provide anecdotal, substantial, strong, very strong, and decisive evidence in favor of the null hypothesis. Thus, unlike frequentist statistics, Bayesian analyses can provide evidence for the null hypothesis.

Descriptive statistics for the MET are provided in Table 2. Distributions are illustrated in Fig. 2. In Tables 9, 10, and 11, respectively, of the Appendix, raw scores are converted to

Table 1 Reliability statistics, including Cronbach's alpha and split-half (odd-even) correlations (Spearman–Brown formula), for scores on the MET

	Melody	Rhythm	Total
Whole sample (N = 523)			
Cronbach's Alpha	.73	.62	.78
Split-half correlation	.71	.68	.78
No music training (n = 189)			
Cronbach's alpha	.59	.56	.68
Split-half correlation	.58	.63	.72
≥ 10 Years of training (n = 98)			
Cronbach's alpha	.65	.64	.77
Split-half correlation	.51	.64	.65

norms for the Melody, Rhythm, and Total measures. Norms include percentiles, Z scores (M = 100, SD = 15), and T scores (M = 50, SD = 10). Table 2 and Fig. 2 confirm that there were no ceiling effects, with no perfect scores and mean levels of performance just under the mid-point between chance and perfect (Melody: 69.3%, Rhythm: 70.1%, Total: 69.7%). Only a small percentage of participants performed below chance levels on either subtest or the complete MET. Distributions departed slightly from normality, ps ≤ .015 (Shapiro–Wilk tests), however, not because of skewness (i.e., the distributions were approximately symmetrical), but because of kurtosis. Specifically, compared to a standard normal distribution, there were relatively few observations in the middle or tails of the curve, but an excess in the shoulders. In general, though, scores were distributed suitably for parametric analyses.

We also observed that listeners had a bias to respond "yes" (same), such that one-sample t tests confirmed that more than 50% (26/52) of responses were "yes" on the Melody subtest (M = 32.94, SD = 5.01), $t(522) = 31.65$, $p < .001$, Cohen's d = 1.38, $BF_{10} > 100$, and on the Rhythm subtest (M = 30.31, SD = 4.78), $t(522) = 20.59$, $p < .001$, Cohen's d = 0.90, $BF_{10} > 100$. In both instances, Bayesian analyses indicated that the observed data provided decisive evidence for this response bias. We expect that on difficult trials, participants often failed to notice when the standard and comparison sequences differed. In any event, all but one of the analyses reported in the present manuscript remained unchanged when *hits* (responding *yes* when the sequences were the same) were adjusted for *false alarms* (responding *yes* when the sequences differed) by calculating *d'* scores, which were almost perfectly correlated with raw scores (rs = .97, .98, and .98, for Melody, Rhythm, and Total scores, respectively, ps < .001, all $BF_{10} > 100$). The one exception was in the analysis of SES, noted below.

Finally, a paired-samples t test found no evidence to suggest that performance accuracy on the Melody and Rhythm subtests differed, $t(522) = 1.86$, $p = .063$, Cohen's d = .081, with the Bayesian counterpart to the same test providing substantial evidence that the observed data were more likely under the null hypothesis, BF_{10} = .273. Melody and Rhythm scores were correlated, however (see Fig. 3), r = .489, N = 523, $p < .001$, as expected, with the observed data providing decisive evidence in favor of an association, $BF_{10} > 100$. Nevertheless, as shown in Fig. 3, the overlapping variance was modest (23.9%), which meant that predictors of performance on the Melody subtest could differ from those of performance on the Rhythm subtest.

Table 2 Descriptive statistics for scores on the MET

	Melody	Rhythm	Total
Whole sample (N = 523)			
Mean	36.05	36.47	72.52
SD	5.36	4.94	8.89
Range	22–49	22–47	50–94
No Music Training (n = 189) N = 523)			
Mean	34.15	35.56	69.71
SD	4.41	4.68	7.48
Range	24–47	24–46	50–88
≥ 10 years of training (n = 98)			
Mean	40.00	38.74	78.74
SD	4.52	4.75	8.13
Range	30–49	25–47	57–94

Maximum scores were 52 for the melody and rhythm subtests (chance = 26), and 104 for the total score (chance = 52)

Demographics

The next set of analyses documented that MET Melody, Rhythm, and Total scores were *not* associated strongly with basic demographic variables including age, gender, and SES. We used standard Pearson correlations for age and SES, and point-biserial correlations for gender. Gender was coded as a binary variable (0 = women, 1 = men). Effect sizes (rs), p-values, and sample sizes are provided in Table 3. There were no associations with age or gender, and Bayesian correlational analyses confirmed that the observed data were at least eight times more likely under the null hypothesis (no association) than the alternative.

SES had very small positive correlations with MET Melody, Rhythm, and Total scores, but when *d'* scores were used the association with Rhythm became non-significant, $p > .1$. Because of our large sample, correlations with raw scores were statistically significant even though the shared variance was miniscule (max: 1.5%). Moreover, Bayesian analysis suggested that evidence in favor of an association was anecdotal at best. In any event, small correlations with SES are evident

Stem and Leaf plots for MET Melody (Upper), Rhythm (Middle), and Total (Lower) scores

Melody

```
2.  23
2.  4444555555
2.  666667777777777
2.  88888888888888999999999999999999
3.  00000000000000000000001111111111111111111111111111
3.  22222222222222222222222233333333333333333333333333333
3.  44444444444444444444444444444444444444555555555555555555555555555555555555555555
3.  66666666666666666666666666667777777777777777777777777777777
3.  88888888888888888888888888888888888888999999999999999999999999999999999999
4.  00000000000000000000000000000001111111111111111111
4.  22222222222222222223333333333333333333
4.  4444444444445555555555
4.  66666667777777
4.  888999999
```

Rhythm

```
2.  23
2.  44555
2.  666666677777777
2.  888888889999999999999
3.  00000000000000001111111111111111111111111
3.  22222222222222222222222222233333333333333333333333333333333
3.  44444444444444444444444444444444444444445555555555555555555555555555555555
3.  66666666666666666666666666666666666666667777777777777777777777777777777777
3.  88888888888888888888888888888888888888999999999999999999999999999999999999999
4.  00000000000000000000000000000000001111111111111111111111111111
4.  22222222222222222222222222233333333333333333
4.  44444444444444444444555555555555555
4.  666666777
```

Total

```
5.  0011
5.  33
5.  4445555
5.  66666677777777
5.  88888889999
6.  0000000011111111
6.  22222222222222233333333333333333
6.  44444444445555555555555555
6.  66666666666666666677777777777777777
6.  88888888888888888888888888889999999999999999999999999999
7.  00000000000000000000000000011111111111111111111
7.  22222222222222222222222222333333333333333333333
7.  44444444444444444444444555555555555555555
7.  66666666666666666777777777777
7.  88888888888888899999999999999999999999999
8.  00000000011111111111111111
8.  22222222222222223333333333333
8.  44444444444444444555555555
8.  6666667777777777777
8.  8888899
9.  00011
9.  22233
9.  444
```

Fig. 2 Stem and leaf plots for MET melody (*upper*), Rhythm (*middle*), and Total (*lower*) scores. *Note:* Each leaf represents a single score, which is added to the stem multiplied by 10. Thus, for Melody, there was one score of 22, one score of 23, four scores of 24, six scores of 25, and so on

for performance on many measures of cognitive ability (grade-point average, IQ; Johnson, McGue, & Iacono, 2007).

Music training

Duration of training

Results from Pearson and Bayesian tests of correlations between MET scores and music training are provided in Table 4. Music training was coded in six different ways: any private training (binary variable: 0 = no training, 1 = some training), duration of private training, any school training (binary variable: 0 = no training, 1 = some training), duration of school training, the sum of years of private and school training (M = 5.0, SD = 7.2), and the square-root transformation of the sum (to reduce positive skew).

As expected, performance on the MET was correlated with all measures of music training except in one instance: the association between Rhythm scores and the binary variable representing school lessons. Bayesian analyses confirmed that the observed associations for all other measures of music training were very strong or decisive. We also tested whether correlations with music training differed in magnitude for Melody scores compared to Rhythm scores, using the test

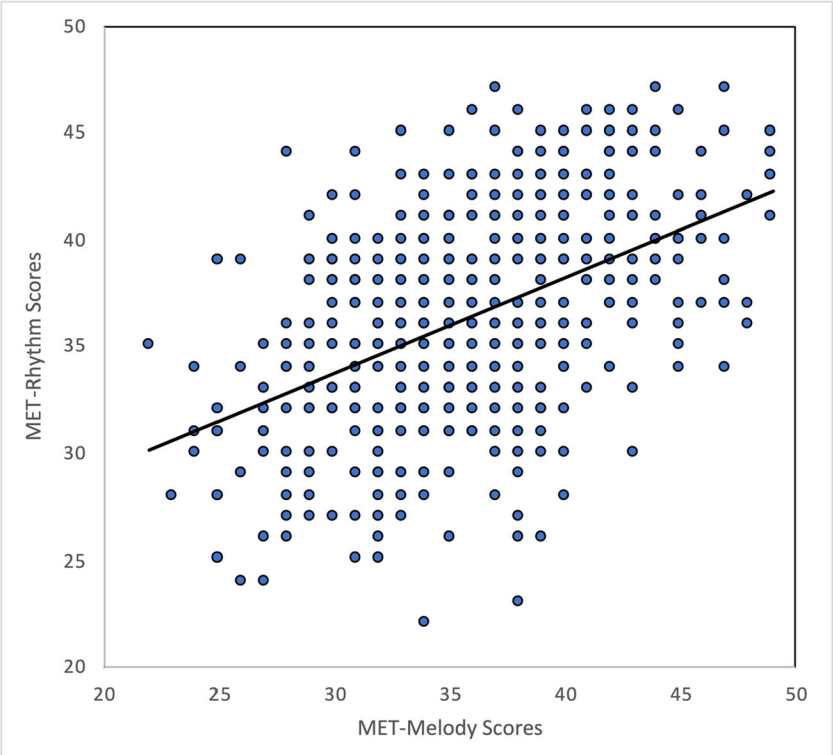

Fig. 3 Scatter plot illustrating MET-rhythm scores (*y*-axis) as a function of MET-melody scores (*x*-axis)

for comparison of correlations from dependent samples.[2] The samples were "dependent" because the same measure of music training (e.g., duration of private lessons) was used to calculate the correlation with Melody scores *and* the correlation with Rhythm scores. As shown in Table 4, Melody scores were better predicted than Rhythm scores for the three measures of music training that included private lessons.

Finally, we asked whether MET scores were better predicted by private compared to school-based music training, again using the test for comparison of correlations from dependent samples. The results are provided in Table 5. For the binary (some-or-none) variables of music training, private training was a better predictor of MET scores than school-based training, and this difference was evident for Melody, Rhythm, and Total scores. Associations with MET scores were similar, however, for *duration* of lessons whether they were taught privately or in school.

As in previous research (Swaminathan et al., 2018; Swaminathan & Schellenberg, 2017, 2018), the *aggregate* measure (square-root duration summed) maximized associations with MET scores for Rhythm scores and for Total scores, at least in terms of absolute magnitude. For Melody scores, the association with the aggregate measure (r = .389) was virtually identical to the association with some private training (r = .394). Subsequent analyses were restricted to the aggregate measure (henceforth, *music training*).

Onset of training

The next set of analyses examined the possibility of associations between performance on the MET and the age at which music training began. Thus, participants with no music training were excluded. According to Penhune (Penhune, 2011; Steele et al., 2013; Watanabe et al., 2007), a sensitive period for effects of music training on development extends to 7

Table 3 Correlations, p values, and Bayes factors (BF_{10}) for associations between MET scores and demographic variables

		Melody	Rhythm	Total
Age	r	.019	.041	.034
	p	.664	.348	.433
	BF_{10}	.060	.085	.074
Gender	r	.057	– .009	.029
	p	.197	.835	.506
	BF_{10}	.126	.056	.069
SES	r	.115	.098	.123
	p	.015	.039	.009
	BF_{10}	1.120	.497	1.775

Note: Age: N = 523, Gender (0 = female, 1 = male): N = 518, SES: N = 448

[2] All comparisons of correlations were conducted with *Psychometrica,* https://www.psychometrica.de/correlation.html).

Table 4 Correlations, p values, and Bayes factors (BF_{10}) for associations between MET scores and music training ($N = 523$). The rightmost column provides comparisons between Melody and Rhythm

		Melody	Rhythm	Total	Melody vs. Rhythm	
Any private lessons	r	.394	.183	.339	z	5.06
	p	< .001	< .001	< .001	p	< .001
	BF_{10}	> 100	> 100	> 100		
Duration-private lessons	r	.314	.172	.285	z	3.34
	p	< .001	< .001	< .001	p	< .001
	BF_{10}	> 100	> 100	> 100		
Any school lessons	r	.138	.069	.121	z	1.57
	p	.002	.117	.005	p	.116
	BF_{10}	8.12	.186	2.59		
Duration-school lessons	r	.222	.168	.227	z	1.25
	p	< .001	< .001	< .001	p	.212
	BF_{10}	> 100	88.6	> 100		
Duration summed	r	.357	.220	.337	z	3.30
	p	< .001	< .001	< .001	p	< .001
	BF_{10}	> 100	> 100	> 100		
Square-root duration summed	r	.389	.228	.361	z	3.88
	p	< .001	< .001	< .001	p	< .001
	BF_{10}	> 100	> 100	> 100		

years of age. We therefore compared participants who began taking lessons by age 7 ($n = 77$) with those who started later in life ($n = 248$).

Independent-samples t tests confirmed that the early starters (M = 39.26, SD = 5.00) performed better than the late starters (M = 36.37, SD = 5.53) on the Melody subtest, $t(323) = 4.09$, $p < .001$, Cohen's $d = .534$, and the Rhythm subtest (early: M = 38.10, SD = 5.10; late: M = 36.68, SD = 4.98), $t(323) = 2.18$ $p = .030$, Cohen's $d = .285$. The early starters (M = 77.36, SD = 8.61) also had higher Total scores than the late starters (M = 73.05, SD = 9.23), $t(323) = 3.64$, $p < .001$, Cohen's $d = .475$. The Bayesian counterpart to an independent-samples t test revealed that the observed data provided decisive and very strong evidence for an early-starter superiority for Melody, $BF_{10} > 100$, and Total, $BF_{10} = 68.4$, scores, respectively. For Rhythm scores, by contrast, the effect was very weak, $BF_{10} = 1.35$.

Early starters (M = 13.88, SD = 9.70) also had more music training compared to late starters (M = 6.10, SD = 5.83), $t(323) = 8.98$, $p < .001$, Cohen's $d = 1.172$, as one might expect, and the difference was decisive according to Bayesian statistics, $BF_{10} > 100$. Thus, we then asked whether the association between age of onset and musical competence was independent of duration of training. We used analysis of covariance (ANCOVA) with onset of training as the independent variable and music training as the covariate. The advantage for early starters disappeared for Melody scores, $F(1, 322) = 1.51$, $p > .2$, Rhythm scores, $F < 1$, and Total scores, $F < 1$. The Bayesian ANCOVAs indicated that the observed data provided substantial evidence for the null hypothesis (i.e., no effect for onset of training) for Rhythm, $BF_{10} = .172$, and Total, $BF_{10} = .227$, scores, and weak evidence for Melody scores, $BF_{10} = .338$.

In the next analysis, we correlated music training with MET scores separately for early and late starters. The results

Table 5 Comparisons of the magnitude of correlations between MET scores and private music training, and MET scores and school-based training ($N = 523$)

	Any lessons				Duration of lessons			
	Private r	School r	z	p	Private r	School r	z	p
Melody	.394	.138	4.72	< .001	.314	.222	1.72	.086
Rhythm	.183	.069	2.00	.046	.172	.168	0.07	.942
Total	.339	.121	3.95	< .001	.285	.227	1.08	.280

are provided in Table 6. Contrary to the plasticity perspective, the correlations were significant and strong for *late* starters, but non-existent for early starters, which raises the possibility that age of onset might moderate the effect of music training. We tested the interaction between music training and onset of training with a general linear model that included both main effects and the two-way interaction. The interaction between music training and onset of training was significant for Melody scores, $F(1, 321) = 5.11$, $p = .025$, $\eta^2 = .014$, $BF_{10} = 1.88$, but not for Rhythm scores, $F < 1$, $p > .5$, $\eta^2 = .001$, $BF_{10} = .240$, and only marginal for Total scores, $F(1, 321) = 2.83$, $p = .093$, $\eta^2 = .008$, $BF_{10} = .658$. In general, then, there was only weak evidence that the magnitude of the association was stronger for late than for early starters.

Finally, we considered only those participants with at least 10 years of music lessons ($n = 95$). Independent-samples t tests revealed no group differences between early and late starters for Melody scores, $t(93) = 0.16$, $p = .877$, Cohen's $d = .032$, $BF_{10} = .218$; Rhythm scores, $t(93) = 0.39$, $p = .696$, Cohen's $d = .081$, $BF_{10} = .231$; or Total scores, $t(93) = 0.31$, $p = .755$, Cohen's $d = .064$, $BF_{10} = .225$; and the Bayes factors indicated that the observed data were substantially more likely under the null than the alternative hypothesis in each instance. We also considered only those participants with at least 10 years of *private* music lessons ($n = 60$). Again, there were no group differences based on onset of training for Melody scores, $t(58) = 0.05$, $p = .960$, Cohen's $d = .013$, $BF_{10} = .265$; Rhythm scores, $t(58) = 0.35$, $p = .731$, Cohen's $d = .090$, $BF_{10} = .278$; or Total scores, $t(58) = 0.22$, $p = .826$, Cohen's $d = .057$, $BF_{10} = .270$; and Bayesian analysis provided substantial evidence for the null hypothesis in each case.

The findings did not change when onset of training was treated as a continuous variable.

General cognitive ability

The next set of analyses asked whether individual differences in performance on the MET could be predicted by individual differences in general cognitive ability as measured by Digit Span Forward, Digit Span Backward, and Raven's test. The results are provided in Table 7. Pearson correlations ranged between .2 and .3, and all were statistically significant and considered decisive or very strong by Bayesian analyses, even though they were modest in size. For all cognitive variables, associations were similar for Melody and Rhythm. In short, performance on the MET varied in tandem with performance on nonmusical measures of cognitive ability.

Table 6 Correlations between music training and MET scores among musically trained participants, reported separately for early and late starters

	Early starters ($N = 77$)			Late starters ($N = 248$)		
	r	p	BF_{10}	r	p	BF_{10}
Melody	.131	.348	.263	.389.	< .001	> 100
Rhythm	.199	.153	.463	.249	.001	19.7
Total	.195	.162	.443	.367	< .001	> 100

Language background

The next set of analyses considered whether knowledge of a tone language would predict performance on the Melody but not on the Rhythm subtest, as it did with a subsample of our participants (Swaminathan et al., 2018) and with participants tested in a different laboratory (Zhang et al., 2020). We initially classified participants into three groups according to their native (first) language: English, a non-tone language other than English, or a tone language. A mixed-design ANOVA with subtest (Melody, Rhythm) as a repeated measure and native language as a between-subjects variable confirmed that native language interacted with subtest, $F(2, 513) = 6.24$, $p = .002$, partial $\eta^2 = .024$. Evidence for the interaction was substantial according to Bayesian statistics, $BF_{10} = 8.31$, which compared the relative likelihood of the observed data with two models: one with only the two main effects (subtest, native language), and another with the two main effects *plus* the two-way interaction.

The significant interaction was followed up with separate one-way ANOVAs for the two MET subtests, with native language as the independent variable. The three groups performed similarly on the Rhythm subtest, $F(2, 513) = 1.95$, $p = .143$, $\eta^2 = .008$, with substantial evidence favoring the null hypothesis, $BF_{10} = .170$. As expected, the groups performed differently on the Melody subtest, $F(2, 513) = 10.71$, $p < .001$, $\eta^2 = .040$, with Tukey's follow-up comparisons showing significantly better performance for the tone-language group than for the other two groups, $ps \leq .001$, who did not differ, $p > .1$. Although the overall effect size was not large, the Bayes factor indicated that the observed data provided decisive evidence for differences in Melody scores as a function of native language, $BF_{10} > 100$.

As noted, many of our participants whose native language was English (or another non-tone language) also spoke a tone language with varying degrees of proficiency. In the next analysis, we considered participants' self-reports of proficiency, from which we formed a continuous measure of tone-language ability, ranging from 0 (no knowledge of a tone language) to 49 (perfect fluency). This measure was correlated positively and decisively with Melody, $r = .215$, $N = 504$, $p < .001$, $BF_{10} > 100$, but not with Rhythm, $r = .034$, $N = 504$, $p = .448$. For Rhythm, the observed data provided decisive evidence favoring the null hypothesis, $BF_{10} = .074$. The interaction between tone language and subtest was re-confirmed by showing that the continuous measure was correlated with

Table 7 Correlations, p values, and Bayes factors (BF_{10}) for associations between MET scores and measures of general cognitive ability

		Melody	Rhythm	Total	Melody vs. Rhythm	
Digit Span Forward	r	.258	.311	.327	z	–1.08
	p	< .001	< .001	< .001	p	.282
	BF_{10}	> 100	> 100	> 100		
Digit Span Backward	r	.296	.353	.372	z	–1.18
	p	< .001	.001	< .001	p	.240
	BF_{10}	> 100	> 100	> 100		
Raven's test	r	.168	.217	.222	z	–1.06
	p	< .001	< .001	< .001	p	.290
	BF_{10}	36.8	> 100	> 100		

The rightmost column provides tests of whether associations differ for the Melody and Rhythm subtests

Note: N = 381 for digit span forward and digit span backward. N = 456 for Raven's test

difference scores (i.e., Melody – Rhythm), $r = .193$, $N = 504$, $p < .001$. The association was small yet decisive according to the Bayes factor, $BF_{10} > 100$. In other words, as proficiency with a tone language improved, Melody scores tended to increase, such that the performance advantage for the Melody subtest over the Rhythm subtest increased as well. These findings did not change when participants from Swaminathan et al. (2018) were excluded.

Multiple regression analyses

In the final set of analyses, we used multiple regression to model MET scores as a function of six variables, each of which had significant simple associations with MET scores: music training, Digit Span Forward, Digit Span Backward, Raven's test, tone-language proficiency (continuous variable), and SES. As noted in the introduction, we had specific hypotheses about music training, Digit Span Forward, and tone-language proficiency, whereas Digit Span Backward and Raven's test controlled for individual differences in nonmusical ability. A Bayesian counterpart to multiple regression considered the same six-predictor model by removing each predictor from the model one at a time to determine whether the observed data were better explained when the predictor was included. The results are provided in Table 8.

The overall model was significant for Melody, Rhythm, and Total scores, but the independent contribution of the individual predictors varied. Melody and Total scores were best predicted by music training, Digit Span Forward, and tone-language proficiency, and the Bayesian analyses provided decisive support for the inclusion of each variable in the model. Rhythm scores were best predicted by music training, Digit Span Forward, and Digit Span Backward, with music training and Digit Span Forward making decisive contributions, and Digit Span Backward making a strong contribution.

Discussion

We used data from a large sample of Canadian undergraduates to compile norms for performance on the MET. We also compared the Melody and Rhythm subtests, and we identified individual-difference variables that predicted performance. The similarity between the Melody and Rhythm subtests in terms of mean levels of performance confirmed that researchers can make direct and meaningful comparisons between subtests. Whereas some individuals will perform similarly on both subtests, others will perform better on one subtest or the other. Researchers can also use our norms to determine how well participants score in relation to our sample. Scores were close to normally distributed for the Total measure and for the Melody and Rhythm subtests, with no evidence of floor or ceiling effects. The one departure from normality was kurtosis.

Associations between MET scores and demographic variables—age, gender, SES—were non-existent or weak. Age had a very restricted range in our sample, however, so the null effect should be taken with a grain of salt. The lack of an association with gender is consistent with current perspectives that the role of women in the history of Western music has taken a backseat to men because of social and cultural reasons, and not because of any gender differences in musical ability (Lumsden, 2010; Wentlent, 2016). Finally, a very weak association with SES was noted, a common finding for many tests of cognitive abilities.

As in previous research with the MET (Wallentin et al., 2010a), music training was associated positively and robustly with test performance, as it has been with other objective measures of musical abilities (Fujii & Schlaug, 2013; Law & Zentner, 2012; Ullén et al., 2014). This association provided evidence of criterion validity for the MET. Nevertheless, the stronger association for Melody than for Rhythm, at least with

Table 8 Summary of multiple-regression analyses predicting MET scores ($N = 363$)

	Melody			Rhythm			Total		
	β	p	BF_{10}	β	p	BF_{10}	β	p	BF_{10}
Music training	.331	< .001	> 100	.173	< .001	75.4	.293	< .001	> 100
Digit span forward	.210	< .001	> 100	.219	< .001	> 100	.246	< .001	> 100
Digit span backward	.084	.126	.532	.166	.004	10.1	.141	.009	4.47
Raven's test	.037	.439	.227	.098	.055	1.17	.076	.110	.567
Tone-language facility	.258	< .001	> 100	.081	.109	.682	.199	< .001	> 100
SES	.085	.065	.889	.078	.108	.685	.093	.039	1.30
R^2	.292	< .001		.210	< .001		.314	< .001	
Adjusted R^2	.280	< .001		.197	< .001		.302	< .001	
$F(6, 353)$	24.241	< .001		15.665	< .001		26.938	< .001	

Bayes factors from Bayesian linear regression are included.

private music lessons, is difficult to explain. One possibility is that it stems from the fact that the chromatic scale, from which the Melody stimuli were constructed, is not a universal characteristic of musical systems. Neither is the use of pitch-based sonorities, as in African drumming music (Jones, 1959). Thus, the skills required to perform well on the Melody subtest may be inherently more cultural (or learned) than those required to perform well on the Rhythm subtest, and therefore more influenced by the learning acquired through formal music lessons. Another possibility is that the pedagogical style used in private and school lessons for teaching music in Canada emphasizes melody over rhythm skills, except when individuals specifically seek training in percussion.

Although associations between musical expertise and music training were strong, age-of-onset of music training was almost completely independent of performance on the MET when duration-of-training was held constant. Bayesian statistics suggested, moreover, that sample size was unlikely to be implicated because evidence in favor of the null hypothesis was substantial, and associations between music training and MET scores were actually significant for late but not for early starters. These findings were unexpected and inconsistent with theory and data reported by other researchers (e.g., Bailey & Penhune, 2012; Penhune, 2011, 2019; Steele et al., 2013; Watanabe et al., 2007). Sensitive periods are evident for other aspects of musical behavior (for review see Trainor, 2005), as they are for language. For example, in the language domain, it is well known that earlier exposure to a second language predicts fluency (Abrahamsson, 2012) and native-like pronunciation (Flege & Fletcher, 1992; Flege, Yeni-Komshian, & Liu, 1999). In the music domain, deprivation studies with animals suggest that exposure to harmonically rich, temporally patterned tones early in life is essential for the proper formation of tonotopic maps and brain circuits for pitch processing (Chang & Merzenich, 2003). Among humans, a younger age of onset is also predictive of acquiring absolute pitch, even though most individuals who begin taking music lessons at an early age do not have absolute pitch (Deutsch, 2013).

The apparent discrepancy between earlier findings and ours may be due to differences in sampling. Previous research on music training and sensitive periods (in samples of individuals *without* absolute pitch) has tended to focus on highly skilled musicians. In these samples, early onset of training predicted differences in brain structure (Baer et al., 2015; Bailey et al., 2014; Steele et al., 2013) and rhythm perception and/or production (Bailey et al., 2014; Bailey & Penhune, 2013; Matthews, Thibodeau, Gunther, & Penhune, 2016; Watanabe et al., 2007). Perhaps early onset predicts musical ability only among musicians who are more skilled than our participants, even those who had 10 or more years of private lessons. Another possibility is that early onset may predict some musical abilities (e.g., rhythm synchronization) but not others (e.g., performance on the MET). A third possibility is that our music-background questionnaire did not give us reliable responses, in contrast to the Musical Experience Questionnaire (Bailey & Penhune, 2010), which was used in the samples of highly skilled musicians. Nevertheless, if onset of training were indeed important but measured inadequately, it seems odd that duration of music training was a robust predictor of musical ability in our large sample, and therefore measured adequately.

Positive correlations between performance on the MET and nonmusical cognitive abilities were in line with predictions. The small but decisive associations were also consistent with Carroll's (1993) three-strata model of intellect, which posits that all abilities are correlated with general ability (*g*, at stratum III) and with each other. Although performance on the MET was correlated with general cognitive skills as measured by Raven's test, Digit Span Forward, and Digit Span Backward, when all three variables were considered simultaneously, Digit Span Forward was the strongest predictor of

MET scores. This result makes sense given the structure of the MET trials, which required the participant to hold the standard sequence in mind while determining whether it was identical to the comparison sequence that followed. A separate, independent contribution of Digit Span Backward was associated with Rhythm subtest performance. Perhaps the wide variation in onset-to-onset durations on test trials required participants to compare different sections of the stimuli in succession, increasing task demands to make them similar to those of Digit Span Backward.

Our finding of a positive association between tone-language experience and melody processing is consistent with earlier findings from behavioral (Bidelman et al., 2013; Zhang et al., 2020) and neuronal (Bidelman et al., 2011; Krishnan, Gandour, Bidelman, & Swaminathan, 2009) studies, and lends itself to a relatively straightforward explanation. When learning a tone language, individuals must attend to differences in pitch and to changes in pitch, which signal a word's semantic meaning. This learning, which begins in infancy (Mattock & Burnham, 2006; Mattock, Molnar, Polka, & Burnham, 2008; Tsao, 2017), appears to enhance the representation and discrimination of pitch in non-linguistic contexts. Although our data are correlational, the reverse causal direction is impossible because of the developmental timeline of native language acquisition. It remains possible, however, that a third, unmeasured variable caused individuals to (1) speak a tone language at home, and (2) be proficient on tests of melody perception and discrimination.

In sum, our results confirm the utility of the MET as an objective index of musical ability. Good performance on the MET is predicted by amount of music training and by individual differences in immediate recall for nonmusical auditory stimuli. Although the Melody and Rhythm subtests are equally difficult, their correlates differ. Scores on the Melody subtest are better explained by private music training compared to those on the Rhythm subtest. Higher Melody scores are also evident among participants who speak a tone language.

Research interest in musical ability measured objectively has grown over recent years, with several reports documenting associations in both adulthood and childhood. For example, in adulthood, researchers have reported links between musical ability and speech perception (Mankel, Barber, & Bidelman, 2020; Mankel & Bidelman, 2018; Swaminathan & Schellenberg, 2017), second-language proficiency (Bhatara, Yeung, & Nazzi, 2015; Roncaglia-Denissen, Roor, Chen, & Sadakata, 2016; Roncaglia-Denissen, Schmidt-Kassow, Heine, Vuust, & Kotz, 2013; Slevc & Miyake, 2006), executive functions (Slevc, Davey, Buschkuehl, & Jaeggi, 2016), short-term memory (Hansen et al., 2013), recognition of vocal emotions (Correia et al., 2020), intelligence (Swaminathan et al., 2017), reading comprehension (Swaminathan et al., 2018), sensitivity to speech rhythms (Magne, Jordan, & Gordon, 2016), faking a foreign accent (Coumel, Christiner, & Reiterer, 2019), personality (Swaminathan & Schellenberg, 2018; Thomas, Silvia, Nusbaum, Beaty, & Hodges, 2016), and working in a creative occupation (Theorell, Madison, & Ullén, 2019). In childhood, links have been reported between musical ability and grammatical ability (Gordon et al., 2015; Lee, Ahn, Holt, & Schellenberg, 2020; Swaminathan & Schellenberg, 2019), phonological processing (Anvari, Trainor, Woodside, & Levy, 2002), and speech perception (Swaminathan & Schellenberg, 2019). The correlates of musical ability identified here need to be considered carefully when: (1) associations between musical ability and other abilities are reported, (2) attempts are made to manipulate musical ability, and (3) the goal is to make causal conclusions about how such ability develops.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-020-01528-8.

Funding Funded by the Natural Sciences and Engineering Research Council of Canada, and the Fundação para a Ciência e a Tecnologia (FCT) em Portugal.

Appendix

Table 9 Raw MET Melody scores converted to percentile, Z-, and T-scores

Raw Score	Percentile	Z-score	T-score
49	99	136	74
48	99	133	72
47	98	131	70
46	97	128	69
45	95	125	67
44	93	122	65
43	91	119	63
42	88	117	61
41	84	114	59
40	80	111	57
39	74	108	56
38	67	105	54
37	60	103	52
36	54	100	50
35	48	97	48
34	40	94	46
33	32	91	44
32	26	89	42
31	21	86	41
30	15	83	39
29	11	80	37
28	8	77	35
27	5	75	33
26	3	72	31
25	2	69	29
24	1	66	28
23	< 1	64	26
22	< 1	61	24

Table 10 Raw MET Rhythm scores converted to percentile, Z-, and T-scores

Raw Score	Percentile	Z-score	T-score
47	99	132	71
46	99	129	69
45	98	126	67
44	95	123	65
43	91	120	63
42	88	117	61
41	83	114	59
40	78	111	57
39	71	108	55
38	63	105	53
37	56	102	51
36	50	99	49
35	42	96	47
34	36	92	45
33	28	89	43
32	22	86	41
31	16	83	39
30	11	80	37
29	8	77	35
28	6	74	33
27	4	71	31
26	3	68	29
25	1	65	27
24	< 1	62	25
23	< 1	59	23
22	< 1	56	21

Table 11 Raw MET Total scores converted to percentile, Z-, and T-scores

Raw Score	Percentile	Z-score	T-score
94	99	136	74
93	99	135	73
92	99	133	72
91	98	131	71
90	98	129	7-
89	98	127	69
88	97	126	67
87	96	124	66
86	94	123	65
85	93	121	64
84	91	119	63
83	87	118	62
82	85	116	61
81	82	114	60
80	79	113	58
79	77	111	57
78	72	109	56
77	69	108	55
76	67	106	54
75	64	104	53
74	60	102	52
73	56	101	51
72	52	99	49
71	47	97	48
70	43	96	47
69	38	94	46
68	33	92	45
67	28	91	44
66	25	89	43
65	22	87	42
64	18	86	40
63	16	84	39
62	13	82	38
61	10	81	37
60	9	79	36
59	7	77	35
58	7	75	34
57	5	74	33
56	4	72	31
55	2	70	30
54	2	69	29
53	1	67	28
52	< 1	65	27
51	< 1	64	26
50	< 1	62	25

References

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition, 34*(2), 187–214. https://doi.org/10.1017/S0272263112000022

Anvari, S. H., Trainor, L. J., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology, 83*(2), 111–130. https://doi.org/10.1016/S0022-0965(02)00124-8

Baer, L. H., Park, M. T., Bailey, J. A., Chakravarty, M. M., Li, K. Z. H., & Penhune, V. B. (2015). Regional cerebellar volumes are related to early musical training and finger tapping performance. *NeuroImage, 105*, 130–139. https://doi.org/10.1016/j.neuroimage.2014.12.076

Bailey, J. A., & Penhune, V. B. (2010). Rhythm synchronization performance and auditory working memory in early- and late-trained musicians. *Experimental Brain Research, 204*, 91–101. https://doi.org/10.1007/s00221-010-2299-y

Bailey, J., & Penhune, V. B. (2012). A sensitive period for musical training: Contributions of age of onset and cognitive abilities. *Annals of the New York Academy of Sciences, 1252*, 163–170. https://doi.org/10.1111/j.1749-6632.2011.06434.x

Bailey, J. A., & Penhune, V. B. (2013). The relationship between the age of onset of musical training and rhythm synchronization performance: Validation of sensitive period effects. *Frontiers in Auditory Cognitive Neuroscience, 7*:227. https://doi.org/10.3389/fnins.2013.00227

Bailey, J. A., Zatorre, R. J., & Penhune, V. B. (2014). Early musical training: Effects on auditory motor integration and grey matter structure in ventral premotor cortex. *Journal of Cognitive Neuroscience, 26*(4), 755–767. https://doi.org/10.1162/jocn_a_00527

Bentley, A. (1966). *Musical ability in children and its measurement*. New York: October House.

Bhatara, A., Yeung, H. H., & Nazzi, T. (2015). Foreign language learning in French speakers is associated with rhythm perception, but not with melody perception. *Journal of Experimental Psychology: Human Perception and Performance, 41*(2), 277–282. https://doi.org/10.1037/a0038736

Bidelman, G. M., & Chung, W.-L. (2015). Tone-language speakers show hemispheric specialization and differential cortical processing of contour and interval cues for pitch. *Neuroscience, 305*, 384–392. https://doi.org/10.1016/j.neuroscience.2015.08.010

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience, 23*(2), 424–434. https://doi.org/10.1162/jocn.2009.21362

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PloS ONE, 8*(4): e60676. https://doi.org/10.1371/journal.pone.0060676

Bidelman, G. M., & Lee, C.-C. (2015). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage, 120*,191–200. https://doi.org/10.1016/j.neuroimage.2015.06.087

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398. https://doi.org/10.1177/0013164498058003002

Butkovic, A., Ullén, F., & Mosing, M. A. (2015). Personality related traits as predictors of music practice: Underlying environmental and genetic influences. *Personality and Individual Differences, 74*, 133–138. https://doi.org/10.1016/j.paid.2014.10.006

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Chang, E. F., & Merzenich, M. M. (2003). Environmental noise retards auditory cortical development. *Science, 300*(5618), 498–502. https://doi.org/10.1126/science.1082163

Conway, A. R. A., Kane, M. J., Bunting, M.F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769–786. https://doi.org/10.3758/BF03196772

Correia, A. I., Castro, S. L., MacGregor, C., Müllensiefen, D., Schellenberg, E. G., & Lima, C. F. (2020). Enhanced recognition of vocal emotions in individuals with naturally good musical abilities. *Emotion*. Advance online publication. https://doi.org/10.1037/emo0000770

Coumel, M., Christiner, M., & Reiterer, S. M. (2019) Second language accent faking ability depends on musical abilities, not on working memory. *Frontiers in Psychology, 10*:257. https://doi.org/10.3389/fpsyg.2019.00257

Deutsch, D. (2013). Absolute pitch. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 141–182). San Diego: Academic Press.

DiLalla, L. (2017). Behavioral genetics. In *Oxford Bibliographies* (September, 2017 ed.). Oxford, UK: Oxford University Press. https://www.oxfordbibliographies.com/view/document/obo-9780199828340/obo-9780199828340-0010.xml

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–703). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.038

Ericsson, K. A., Rampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

Flege, J., & Fletcher, K. (1992). Talker and listener effects on the perception of degree of foreign accent. *Journal of the Acoustical Society of America, 91*(1), 370–389. https://doi.org/10.1121/1.402780

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

Fujii, S., & Schlaug, G. (2013) The Harvard Beat Assessment Test (H-BAT): A battery for assessing beat perception and production and their dissociation. *Frontiers in Human Neuroscience, 7*:771. https://doi.org/10.3389/fnhum.2013.00771

Gordon, E. (1965). *Musical aptitude profile: Manual*. Boston: Houghton Mifflin.

Gordon, E. E. (1979). *Primary measures of music audiation* (K-Grade 3). Chicago: GIA Publications.

Gordon, E. E. (1982). *Intermediate measures of music audiation* (Grade 1–6). Chicago: GIA Publications.

Gordon, E. E. (1989). *Advanced measures of music audiation* (Grade 7–Adult). Chicago: GIA Publications.

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., McAuley, J. D. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Developmental Science, 18*(4), 635–644. https://doi.org/10.1111/desc.12230

Graham, G. (2019). Behaviorism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 ed.). Stanford University, https://plato.stanford.edu/archives/spr2019/entries/behaviorism

Hambrick, D. Z., Macnamara, B. N., Campitelli, G., Ullén, F., & Mosing, M. A. (2016). Beyond born versus made: A new look at expertise. *Psychology of Learning and Motivation, 64*, 1–55. https://doi.org/10.1016/bs.plm.2015.09.001

Hambrick, D. Z., & Tucker-Drob, E. (2015). The genetics of music accomplishment: Evidence for gene-environment correlation and interaction. *Psychonomic Bulletin & Review*, 22, 112–120. https://doi.org/10.3758/s13423-014-0671-9.

Hansen, M., Wallentin, M., & Vuust, P. (2013). Working memory and musical competence of musicians and non-musicians. *Psychology of Music, 41*(6), 779–793. https://doi.org/10.1177/0305735612452186

Herholz, S. C., & Zatorre, R. J. (2012). Musical training as a framework for brain plasticity: Behavior, function, and structure. *Neuron, 76*(3), 486–502. https://doi.org/10.1016/j.neuron.2012.10.011

Howe, M. J. A., Davidson, J. W., & Sloboda, J. A. (1998). Innate talents: Reality or myth? *Behavioral and Brain Sciences, 21*(3), 399–407. https://doi.org/10.1017/S0140525X9800123X

Hutka, S., Bidelman, G. M., & Moreno, S. (2015). Pitch expertise is not created equal: Cross-domain effects of music and tone language experience on neural and behavioural discrimination of speech and music. *Neuropsychologia, 71*, 52–63. https://doi.org/10.1016/j.neuropsychologia.2015.03.019

Jarosz, A., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving, 7*(1), 2–9. https://doi.org/10.7771/1932-6246.1167

JASP Team (2019). *JASP* (Version 0.10.2) [Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Johnson, E., McGue, M., & Iacono, W. G. (2007). Socioeconomic status and school grades: Placing their association in broader context in a sample of biological and adoptive families. *Intelligence, 35*(6), 526–541. https://doi.org/10.1016/j.intell.2006.09.006

Jones, A. M. (1959). *Studies in African music*. Oxford, UK: Oxford University Press.

Krishnan, A., Gandour, J. T., Bidelman, G. M., & Swaminathan, J. (2009). Experience-dependent neural representation of dynamic pitch in the brainstem. *NeuroReport, 20*(4), 408–413. https://doi.org/10.1097/WNR.0b013e3283263000

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS ONE, 7*(12), e52508. https://doi.org/10.1371/journal.pone.0052508

Lee, Y. S., Ahn, A., Holt, R. F., & Schellenberg, E. G. (2020). Rhythm and syntax processing in school-age children. *Developmental Psychology, 56*(9), 1632–1641. https://doi.org/10.1037/dev0000969

Lumsden, R. (2010). Women's leadership in Western music since 1800. In K. O'Connor (Ed.), *Gender and women's leadership: A reference handbook* (pp. 917–925). Thousand Oaks, CA: Sage.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science, 25*(8), 1608–1618. https://doi.org/10.1177/0956797614535810

Macnamara, B. N., Moreau, D., & Hambrick, D. Z. (2016). The relationship between deliberate practice and performance in sports: A meta-analysis. *Perspectives on Psychological Science, 11*(3), 333–350. https://doi.org/10.1177/1745691616635591

Magne, C., Jordan, D. K., Gordon, R. L. (2016) Speech rhythm sensitivity and musical aptitude: ERPs and individual differences. *Brain and Language, 153–154*, 13–19. https://doi.org/10.1016/j.bandl.2016.01.001

Mankel, K., Barber, J., & Bidelman, G. M. (2020). Auditory categorical processing for speech is modulated by inherent musical listening skills. *NeuroReport, 31*, 162–166. https://doi.org/10.1097/WNR.0000000000001369

Mankel, K., & Bidelman, G. M. (2018). Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proceedings of the National Academy of Sciences of the United States of America, 115*(51), 13129–13134. https://doi.org/10.1073/pnas.1811793115

Matthews, T. E., Thibodeau, J. N. L., Gunther, B. P., & Penhune, V. B. (2016) The impact of instrument-specific musical training on rhythm perception and production. *Frontiers in Psychology, 7*:69. https://doi.org/10.3389/fpsyg.2016.00069

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy, 10*(3), 241–265. https://doi.org/10.1207/s15327078in1003_3

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition, 106*(3), 1367–1381. https://doi.org/10.1016/j.cognition.2007.07.002

Meinz, E. J., & Hambrick, D. Z. (2010). Deliberate practice is necessary but not sufficient to explain individual differences in piano sight-reading skill: The role of WMC. *Psychological Science, 21*(7), 914–919. https://doi.org/10.1177/0956797610373933.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullén, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science, 25*(9), 1795–1803. https://doi.org/10.1177/0956797614541990

Mosing, M. A., Madison, G., Pedersen, N. L., & Ullén, F. (2016). Investigating cognitive transfer within the framework of music practice: Genetic pleiotropy rather than causality. *Developmental Science, 19*(3), 504–512. https://doi.org/10.1111/desc.12306

Mosing, M. A., & Ullén, F. (2018). Genetic influences on musical specialization: A twin study on choice of instrument and music genre. *Annals of the New York Academy of Sciences, 1423*, 427–434. https://doi.org/10.1111/nyas.13626

Müllensiefen, D., Gingras, B., Musil, J., Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS ONE, 9*(2), e89642. https://doi.org/10.1037/t42817-000

Penhune, V. B. (2011). Sensitive periods in human development: Evidence from musical training. *Cortex, 47*(9), 1126–1137. https://doi.org/10.1016/j.cortex.2011.05.010

Penhune, V. B. (2019). Music training and brain structure: The causes and consequences of training. In M. H. Thaut & D. A. Hodges (Eds.), *The Oxford handbook of music and the brain* (pp. 419–438). Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198804123.013.17

Penhune, V. B. (2020). A gene-maturation-environment model for understanding sensitive period effects in musical training. *Current Opinion in Behavioral Sciences, 36*, 13–22. https://doi.org/10.1016/j.cobeha.2020.05.011

Penhune, V., & de Villers-Sidani, E. (2014). Time for new thinking about sensitive periods. *Frontiers in Systems Neuroscience. 8*:55. https://doi.org/10.3389/fnsys.2014.00055

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders. *Annals of the New York Academy of Sciences, 999*, 58–75. https://doi.org/10.1196/annals.1284.006

Piro, J. M., & Oritz, C. (2009). The effect of piano lessons on the vocabulary and verbal sequencing skills of primary grade students. *Psychology of Music, 37*(3), 325–347. https://doi.org/10.1177/0305735608097248

Raven, J. C. (1965). *Advanced Progressive Matrices, Sets I and II*. Toronto: Psychological Corporation.

Roncaglia-Denissen, M. P., Roor, D. A., Chen, A., & Sadakata, M. (2016) The enhanced musical rhythmic perception in second language learners. *Frontiers in Human Neuroscience, 10*:288. https://doi.org/10.3389/fnhum.2016.00288

Roncaglia-Denissen, M.P., Schmidt-Kassow, M., Heine, A., Vuust, P., & Kotz, S. A. (2013). Enhanced musical rhythmic perception in Turkish early and late learners of German. *Frontiers in Psychology, 4*:645. https://doi.org/10.3389/fpsyg.2013.00645

Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. *Journal of Educational Psychology, 98*(2), 457–468. https://doi.org/10.1037/0022-0663.98.2.457

Schellenberg, E. G. (2020). Music training, individual differences, and plasticity. In M. S. C. Thomas, D. Mareschal, & I. Dumontheil (Eds.), *Educational neuroscience: Development across the lifespan* (pp. 413–439). New York: Routledge. https://doi.org/10.4324/9781003016830

Seashore, C. (1919). *The psychology of musical talent*. New York, NY: Holt.

Seashore, C. E., Saetveit, J. G., & Lewis, D. (1960). *Seashore measures of musical talent* (rev. ed.). New York: Psychological Corporation.

Skinner, B. F. (1976). *About behaviorism*. New York: Random House.

Slevc, L. R., Davey, N. S., Buschkuehl, M., & Jaeggi, S. M. (2016). Tuning the mind: Exploring the connections between musical ability and executive functions. *Cognition, 152*, 199–211. https://doi.org/10.1016/j.cognition.2016.03.017

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency. *Psychological Science, 17*, 675–681. https://doi.org/10.1111/j.1467-9280.2006.01765.x

Steele, C. J., Bailey, J. A., Zatorre, R. J., & Penhune, V. B. (2013). Early musical training and white-matter plasticity in the corpus callosum: Evidence for a sensitive period. *Journal of Neuroscience, 33*(3), 1282–1290. https://doi.org/10.1523/JNEUROSCI.3578-12.2013

Steele, C. J., & Zatorre, R. J. (2018). Practice makes plasticity. *Nature Neuroscience, 21*, 1645–1650. https://doi.org/10.1038/s41593-018-0280-4

Stevens, C. J., Keller, P. E., & Tyler, M. D. (2013). Tonal language background and detecting pitch contour in spoken and musical items. *Psychology of Music, 41*(1), 59–74. https://doi.org/10.1177/0305735611415749

Swaminathan, S., & Schellenberg, E.G. (2017). Musical competence and phoneme perception in a foreign language. *Psychonomic Bulletin & Review, 24*(6), 1929–1934. https://doi.org/10.3758/s13423-017-1244-5

Swaminathan, S., & Schellenberg, E.G. (2018). Musical competence is predicted by music training, cognitive abilities, and personality. *Scientific Reports, 8*, 9223. https://doi.org/10.1038/s41598-018-27571-2

Swaminathan, S., & Schellenberg, E. G. (2019). Music training and cognitive abilities: Associations, causes, and consequences. In M. H. Thaut & D. A. Hodges (Eds.), *The Oxford handbook of music and the brain* (pp. 645–670). Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198804123.013.26

Swaminathan, S., Schellenberg, E. G., & Khalil, S. (2017). Revisiting the association between music lessons and intelligence: Training effects or music aptitude? *Intelligence, 62*, 119–124. https://doi.org/10.1016/j.intell.2017.03.005

Swaminathan, S., Schellenberg, E. G., & Venkatesan, K. (2018). Explaining the association between music training and reading in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(6), 992–999. https://doi.org/10.1037/xlm0000493

Thagard, P. (2019). Cognitive science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Palo Alto: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/cognitive-science

Theorell, T., Madison, G., & Ullén, F. (2019). Associations between musical aptitude, alexithymia, and working in a creative occupation. *Psychology of Aesthetics, Creativity, and the Arts, 13*(1), 49–57. https://doi.org/10.1037/aca0000158

Thomas, K. S., Silvia, P. J., Nusbaum, E. C., Beaty, R. E., & Hodges, D. A. (2016). Openness to experience and auditory discrimination ability in music: An investment approach. *Psychology of Music, 44*(4), 792–801. https://doi.org/10.1177/0305735615592013

Trainor, L. J. (2005). Are there critical periods for musical development? *Developmental Psychobiology, 46*(3), 262–278. https://doi.org/10.1002/dev.20059

Tsao, F.-M. (2017). Perceptual improvement of lexical tone in infants: Effects of tone language experience. *Frontiers in Psychology, 8*:558. https://doi.org/10.3389/fpsyg.2017.00558

Ullén, F., Hambrick, D. Z., & Mosing, M. A. (2016). Rethinking expertise: A multifactorial gene–environment interaction model of expert performance. *Psychological Bulletin, 142*(4), 427–446. https://doi.org/10.1037/bul0000033

Ullén, F., Mosing, M. A., Holm, L., Eriksson, H., & Madison, G. (2014). Psychometric properties and heritability of a new online test for musicality, the Swedish Musical Discrimination Test. *Personality and Individual Differences, 63*, 87–93. https://doi.org/10.1016/j.paid.2014.01.057

Vaquero, L., Rousseau, P.-N., Vozian, D., Klein, D., & Penhune, V. (2020). What you learn & when you learn it: Impact of early bilingual & music experience on the structural characteristics of auditory-motor pathways. *NeuroImage, 213*, 116689. https://doi.org/10.1016/j.neuroimage.2020.116689

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010a). The Musical Ear Test: A new reliable test for measuring musical competence. *Learning and Individual Differences, 20*(3), 188–196. https://doi.org/10.1016/j.lindif.2010.02.004

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010b). Corrigendum to "The Musical Ear Test, a new reliable test for measuring musical competence" [Learning and Individual Differences Volume 20 (3) 188–196]. *Learning and Individual Differences, 20*, 705. https://doi.org/10.1016/j.lindif.2010.10.001

Watanabe, D., Savion-Lemieux, T., & Penhune, V. B. (2007). The effect of early musical training on adult motor performance: Evidence for a sensitive period in motor learning. *Experimental Brain Research, 176*(2), 332–340. https://doi.org/10.1007/s00221-006-0619-z

Wentlent, A. (2016). *The women of Western music: Hildegard to Ella.* Van Nuys, CA: Alfred Music.

Wing, H. D. (1962). A revision of the Wing Musical Aptitude Test. *Journal of Research in Music Education, 10*(1), 39–46. https://doi.org/10.2307/3343909

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*(4), 420–422. https://doi.org/10.1038/nn1872

Zhang, L., Xie, S. Li, Y., Shu, H., & Zhang, Y. (2020). Perception of musical melody and rhythm as influenced by native language experience. *Journal of the Acoustical Society of America, 147*(5), EL385-EL390. https://doi.org/10.1121/10.0001179

Open Practices Statements

The data for the study are available in Supplementary Information. The study was not preregistered.