Behavior Research Methods (2022) 54:955–969
https://doi.org/10.3758/s13428-021-01641-2

Can musical ability be tested online?

Ana Isabel Correia[1] · Margherita Vincenzi[1,2] · Patrícia Vanzella[3] · Ana P. Pinheiro[4] · César F. Lima[1,5] · E. Glenn Schellenberg[1,6]

Accepted: 29 May 2021 / Published online: 11 August 2021

Abstract

We sought to determine whether an objective test of musical ability could be successfully administered online. A sample of 754 participants was tested with an online version of the Musical Ear Test (MET), which had Melody and Rhythm subtests. Both subtests had 52 trials, each of which required participants to determine whether standard and comparison auditory sequences were identical. The testing session also included the Goldsmiths Musical Sophistication Index (Gold-MSI), a test of general cognitive ability, and self-report questionnaires that measured basic demographics (age, education, gender), mind-wandering, and personality. Approximately 20% of the participants were excluded for incomplete responding or failing to finish the testing session. For the final sample ($N = 608$), findings were similar to those from in-person testing in many respects: (1) the internal reliability of the MET was maintained, (2) construct validity was confirmed by strong associations with Gold-MSI scores, (3) correlations with other measures (e.g., openness to experience, cognitive ability, mind-wandering) were as predicted, (4) mean levels of performance were similar for individuals with no music training, and (5) musical sophistication was a better predictor of performance on the Melody than on the Rhythm subtest. In sum, online administration of the MET proved to be a reliable and valid way to measure musical ability.

Keywords Music · Ability · Expertise · Training · Melody · Rhythm

For most of us, the Internet is part of everyday life. Over half of the world's population (51%) now uses the Internet, and this proportion is even higher for young people (69%), especially those living in developed countries (98%; International Telecommunication Union, 2020). The COVID-19 pandemic increased the amount of time people spend on the Internet while restricting in-person contact, making online testing an attractive option for psychological research. Even before the pandemic, online methods were increasingly used as an alternative to in-person research conducted in the laboratory (e.g., Chetverikov & Upravitelev, 2015; Houben & Wiers, 2008; Milne et al., 2020; Smith & Leigh, 1997; Taherbhai et al., 2012), while the emergence of a number of online platforms provided new tools for recruitment and testing (e.g., Gosling & Mason, 2015; Grootswagers, 2020).

Although there are legitimate concerns about online testing, such as lack of control over characteristics of the samples and testing contexts (e.g., Birnbaum, 2004; Krantz & Dalal, 2000), online studies have several features that make them equivalent or even superior to in-person testing (e.g., Casler et al., 2013; Dandurand et al., 2008; Gosling et al., 2004). First, data quality can be similar, in the sense that the findings are similar. Second, Internet samples can be more diverse and representative of the general population in terms of age, gender, and socioeconomic status, particularly when compared to samples comprised solely of college students registered in

César F. Lima and E. Glenn Schellenberg are joint last authors

✉ César F. Lima
cesar.lima@iscte-iul.pt

1 Centro de Investigação e Intervenção Social (CIS-IUL), Instituto Universitário de Lisboa (ISCTE-IUL), Av.ª das Forças Armadas, 1649-026 Lisboa, Portugal

2 Department of General Psychology, University of Padova, Padova, Italy

3 Center for Mathematics, Computing, and Cognition, Universidade Federal do ABC, Santo Andre, Brazil

4 CICPSI, Faculdade de Psicologia, Universidade de Lisboa, Lisboa, Portugal

5 Institute of Cognitive Neuroscience, University College London, London, UK

6 Department of Psychology, University of Toronto Mississauga, Mississauga, Canada

introductory psychology courses. Third, access to relatively rare target audiences, such as musicians, tends to be easier. Fourth, participants may feel more comfortable and act more naturally at home than when they come to a laboratory. Fifth, building an online experiment, recruiting participants, and collecting data can be more efficient in terms of time and costs, especially when responses are scored and recorded automatically on the hosting platform. Finally, online experiments are not limited to the space and time constraints of a laboratory.

Despite these benefits, online testing needs specific exclusion criteria, careful experimental designs that maximize control (e.g., Gosling et al., 2004), and appropriate motivational strategies (e.g., promising feedback at the end) to improve the likelihood that participants complete the whole experiment. Auditory research, and temporally based experimental tasks in general, can be particularly challenging, because compared to the laboratory, online testing occurs in contexts that are more variable and uncontrolled in terms of extraneous sounds, technical aspects of stimulus presentation, and potential interruptions (e.g., Milne et al., 2020). Although this variability can be reduced by asking participants to follow specific instructions (e.g., to wear headphones), experimental control remains limited.

How similar are the findings from in-person and online experiments? Positive results come from an online study about reinforcement learning (Nussenbaum et al., 2020), which replicated a main effect of age that was reported in an earlier in-person study (Decker et al., 2016). In other developmental research, online data replicated a mediating role for abstract reasoning ability in the link between age and model-based learning (Chierchia et al., 2019). In non-developmental research, Houben and Wiers (2008) found that an implicit association test was effective at identifying alcohol-related associations whether it was administered online or in person.

Although there is substantial evidence that simple tasks can be reliably adapted for online testing, an open question is whether longer and more cognitively demanding tasks can be similarly adapted. In one instance, Dandurand et al. (2008) adapted a complex problem-solving task (from Dandurand et al., 2004) for online testing. Across platforms, participants' performance was better when they observed or read instructions on how to solve the problem successfully, compared to when they were simply given feedback on their decisions. Nevertheless, online participants were less accurate in general than in-person participants, even though the testing format did not influence the main effect of the learning manipulation (i.e., no interaction).

In the present investigation, we used the platform *Gorilla* (http://www.gorilla.sc/; Anwyl-Irvine et al., 2020) to create an online version of an objective measure of musical ability—the Musical Ear Test (MET). The MET is a listening test that has documented reliability and validity (Swaminathan et al., 2021; Wallentin et al., 2010a, 2010b). It is designed in the tradition of musical *aptitude* (i.e., natural musical ability) tests, with two subtests, Melody and Rhythm, both of which require participants to determine, on multiple trials, whether two auditory sequences (a standard followed by a comparison) are identical. Musical aptitude tests, dating back to the early twentieth century (Bentley, 1966; Gordon, 1965; Seashore, 1919; Seashore et al., 1960; Wing, 1962), were designed to identify whether musically untrained individuals (primarily children) are likely to benefit from music lessons, based on the view that people with little natural ability would be unlikely to benefit in this regard. These older tests, as well as more recent tests of musical ability (Asztalos & Csapó, 2014; Fujii & Schlaug, 2013; Law & Zentner, 2012; Peretz et al., 2003, 2013; Ullén et al., 2014; Zentner & Strauss, 2017), all require same–different comparisons of two auditory events that differ in pitch (e.g., melody) or time (e.g., rhythm), or along other dimensions such as timbre and amplitude. In other words, the tests rely on core musical skills, specifically auditory short-term (working) memory and perceptual discrimination. As a broad phenotype, musical ability incorporates many other aspects of behavior (e.g., expert levels of performance, long-term memory for melodies) that are dependent on learning and practice. The goal of tests such as the MET is to measure musical ability in the absence of any formal training, and to do so objectively and quickly.

We also used *Gorilla* to run the entire testing session, which included measures of general cognitive ability and personality, and to create an online version of a self-report measure of musical behavior and expertise—the Goldsmiths Musical Sophistication Index (Gold-MSI; Lima et al., 2020; Müllensiefen, et al., 2014). The Gold-MSI served as our principal measure of construct validity. Virtually all developers of tests of musical ability report positive correlations with musical expertise as a means of documenting a test's validity (Asztalos & Csapó, 2014; Law & Zentner, 2012; Wallentin et al., 2010a; Zentner & Strauss, 2017; Ullén et al., 2014).

We compared response patterns from our online sample with previous studies that had large samples of participants: Swaminathan et al. (2021, $N = 523$) for the MET, and Lima et al. (2020, $N = 408$) for the Gold-MSI. Specifically, we compared the present sample with these comparison samples in terms of their psychometric characteristics, including internal reliability, construct validity, correlations between subtests, and correlations between musical ability and musical sophistication. We also tested for associations with demographic variables, cognitive ability, and personality, because previous studies have shown robust associations with these variables (e.g., Cooper, 2019; Greenberg et al., 2015; Kuckelkorn et al., 2021; Lima et al., 2020; Moreno et al., 2011; Swaminathan et al., 2021). Absolute levels of performance on our measures could vary across samples depending on the degree to which they differ in music training, age, cognitive ability,

personality, education, and so on. In terms of age and education, Lima et al. tested Portuguese individuals from the general population who varied widely, whereas Swaminathan et al. tested Canadian undergraduates who varied minimally.

Because the Gold-MSI has a history of online *and* in-person testing (Correia et al., 2020; Greenberg et al., 2015; Lima et al., 2020; Müllensiefen et al., 2014; Schaal et al., 2015), we predicted that results from our online version of the test would be similar to those from the paper-and-pencil administration of Lima et al. (2020), with similar psychometric properties. We were less certain of the outcome with the online version of the MET, primarily because technological requirements were much greater for an objective listening test, which required participants to determine, on each of 104 trials, whether two auditory sequences were identical.

In short, our main objective was to determine whether the MET could be successfully administered online. Evidence of *success* required that the test's internal reliability would not be compromised by online administration, that performance would correlate positively with musical expertise, and that musical ability would have positive associations with general cognitive ability. Moreover, musical expertise should be a better predictor of scores for the Melody subtest of the MET than for the Rhythm subtest, as is the case with in-person testing (Swaminathan et al., 2021). Other findings from previous research (Swaminathan & Schellenberg, 2018; Butkovic et al., 2015) indicated that the online test's success would be further supported by a positive correlation with scores on one (and only one) dimension from the Big Five model of personality (McCrae & Costa, 1987; McCrae & John, 1992): openness to experience.

More novel aspects of the present study included our prediction that mind-wandering would be associated negatively with performance on the MET, because the MET required participants to concentrate for 18 min. One might also expect lower levels of mind-wandering among individuals who have taken music lessons for a longer period of time, because learning to play music requires much time, effort, and focus. Our use of the Gold-MSI as a measure of musical expertise allowed us to explore whether aspects of musical expertise other than training were predictive of performance, and whether their predictive power would vary across subtests. Previous studies of musical ability restricted tests of construct validity to associations with musicianship status, amount of daily practice, duration of music training, or involvement in professional music-related activities (Law & Zentner, 2012; Swaminathan et al., 2021; Ullén et al., 2014; Wallentin et al., 2010a). The Gold-MSI allowed us to examine whether musical ability would also be associated with active engagement with music, emotional responding to music, and self-reports of singing and perceptual abilities. Such associations would confirm that the narrow range of abilities tested by the MET is predictive of a much broader range of musical abilities.

Method

Participants

A total of 754 participants were tested originally. We subsequently excluded participants who did not complete the MET ($n = 100$) or failed to respond on several trials on either the Melody or the Rhythm subtest, which we defined as more than 10 trials in total ($n = 39$) or more than 5 in a row ($n = 7$). The final sample included 608 participants (361 female, 243 male, 4 unreported) between 18 and 88 years of age ($M = 34.2$, $SD = 15.1$). Most had completed high school ($n = 207$) or had a university degree (bachelor's, $n = 108$, master's, $n = 191$, Ph.D., $n = 58$). Only three participants had less than 10 years of education. Education data were missing for 41 participants.

Participants were recruited primarily through snowball sampling and social media posts, which read: *Do you like music? Do you know anyone who does? We are running an online study on personality and musical abilities. We are looking for listeners with all kinds of musical backgrounds.* A subsample of undergraduate students was recruited via email and received partial course credit for their participation. The experiment was available in four languages, and participants were instructed to complete it in their native language (Italian, $n = 288$; European Portuguese, $n = 153$; Brazilian Portuguese, $n = 123$; English, $n = 44$). Informed consent was collected from all participants, and ethical approval for the study protocol was obtained from the local ethics committee at ISCTE-IUL (reference 07/2021).

Participants varied widely in terms of music training. Half had no history of music lessons ($n = 151$) or a maximum of 2 years ($n = 133$), but 156 had 10 years or more. The training included private lessons ($n = 123$), or classes taught at university ($n = 122$) or in musical academies or conservatories ($n = 84$). Others ($n = 85$) were self-taught. On average, participants with music lessons started their training at the age of 11.4 years ($SD = 7.1$; range: 2–56). The relatively high proportion of participants with extensive backgrounds in music was presumed to stem from their personal interest in the study.

Measures

All tasks and questionnaires, created originally in English, were adapted for online testing using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Validated translations of the measures (e.g., the Big Five Inventory in European-Portuguese and Italian) were used when available. When a task or questionnaire was not available for our target languages, instructions and items were translated by bilinguals who were native speakers and also fluent in English.

Online versions of the MET and the Gold-MSI are available on Gorilla for other researchers to use (https://app.gorilla.sc/openmaterials/218554).

Objective behavioral tests

Musical ability An online version of the Musical Ear Test (MET; Wallentin et al., 2010a) was used to evaluate music perception abilities. We attempted to make the online experience as similar as possible to in-person testing, when the test is installed on a personal computer in the laboratory, and participants listen to stimuli over headphones and record their responses on an answer sheet. As in the original version, the online MET had two subtests, Melody and Rhythm (in that order), each of which had 52 trials. On each trial, participants listened to two short musical excerpts (a standard followed by a comparison) and made a yes/no judgment about whether the comparison was the same as the standard. On both subtests, half of the trials were *same* and half were *different*. The stimuli and order of presentation were the same as in the original test. All musical excerpts had the same metrical structure (4/4 time) and tempo (100 beats per minute). A lower-amplitude metronome sound indicated the underlying beat. Each subtest was preceded by two practice trials (one *same*, one *different*). Feedback was provided for practice trials but not for test trials. Detailed descriptions of MET stimuli are provided in Swaminathan et al. (2021).

In the original test, all instructions and trials are presented via an 18-min digital audio file, with task instructions and the number of each trial provided by a male speaker. Trials are not self-paced. Rather, participants are given a brief window after each trial (1500 ms for melodic trials, 1659 to 3230 ms for rhythmic trials) to respond by checking *yes* or *no* on a response sheet. In our online adaptation of the MET, instructions and trial numbers were converted to text that participants read. The actual stimuli from each trial were digitally copied from the original audio file and the duration of the interstimulus intervals was preserved, such that the total duration (approximately 20 min) of the MET was identical to the in-person version. The trial number and the question (e.g., *Are the melodic phrases identical?)* were visible on the screen from the beginning of each trial until the participant responded. Immediately after the audio stimulus ended, two buttons—labeled *Yes* and *No*—appeared, and participants had a few moments to respond by clicking the appropriate button. Examples of MET stimuli are illustrated in musical notation in Fig. 1.

To enhance the online testing experience, we provided a progress bar at the bottom of the screen throughout both subtests, such that participants could monitor where they were in relation to the beginning and end of the subtest. We also provided feedback at the end of the test about the participant's performance, which was calculated as the total number of correct responses on the Melody and Rhythm subtests. For statistical analyses, a Total score was also calculated as the sum.

General cognitive ability Our measure of general cognitive ability (hereafter *cognitive ability*) was the Matrix Reasoning Item Bank (MaRs-IB; Chierchia et al., 2019), an online test of abstract (nonverbal) reasoning modeled after Raven's Advanced Progressive Matrices (Raven, 1965). On each of 80 trials, a 3 × 3 matrix was presented on the computer screen. Eight of nine cells contained abstract shapes, but the ninth (bottom-right) cell was always empty. Participants' task was to complete the matrix by choosing one of four alternatives. Two examples are provided in Fig. 2. Associations among shapes could vary on a single dimension for the simplest trials (e.g., color), but on up to four dimensions (e.g., color, size, shape, and location) for more difficult trials.

On each trial, before the matrix was presented, a 500-ms fixation cross appeared in the middle of the screen, followed by a 100-ms white screen. Participants then had up to 30 s to look at the matrix and select a response. The trial ended earlier if participants responded. If no response was provided after 25 s, a clock appeared and indicated the time remaining.

The order of the trials was the same for all participants. The first five items were relatively easy so as to familiarize participants with the task. Although the duration of the entire task was fixed at 8 min, participants were not informed of the task duration or the number of trials—only that they had up to 30 s to complete each trial. If they completed the 80 trials in less than 8 min, the trials were presented again in the same order, but responses from the second round were not considered in calculating scores. Scores were calculated as the proportion of the total number of responses given by the participant that were correct. For the statistical analyses, proportions were logit-transformed.

Questionnaires

Musical expertise Our principal measure for tests of construct validity was the Gold-MSI (Müllensiefen et al., 2014), a self-report questionnaire of musical expertise and behavior. The Gold-MSI has 38 items that evaluate different behaviors related to music (e.g., *I spend a lot of my free time doing music-related activities*). Although the items are mixed in terms of order of presentation, for scoring purposes they are grouped to form five subtests: Active Engagement (9 items), Perceptual Abilities (9 items), Music Training (7 items), Singing Abilities (7 items), and Emotions (6 items). A General Musical Sophistication factor is also calculated from 18 items that are representative of the five subtests. For the first 31 items, participants judge how much they agree with each statement on a seven-point rating scale (1 = *completely disagree*, 7 = *completely agree*). For the final seven items, participants

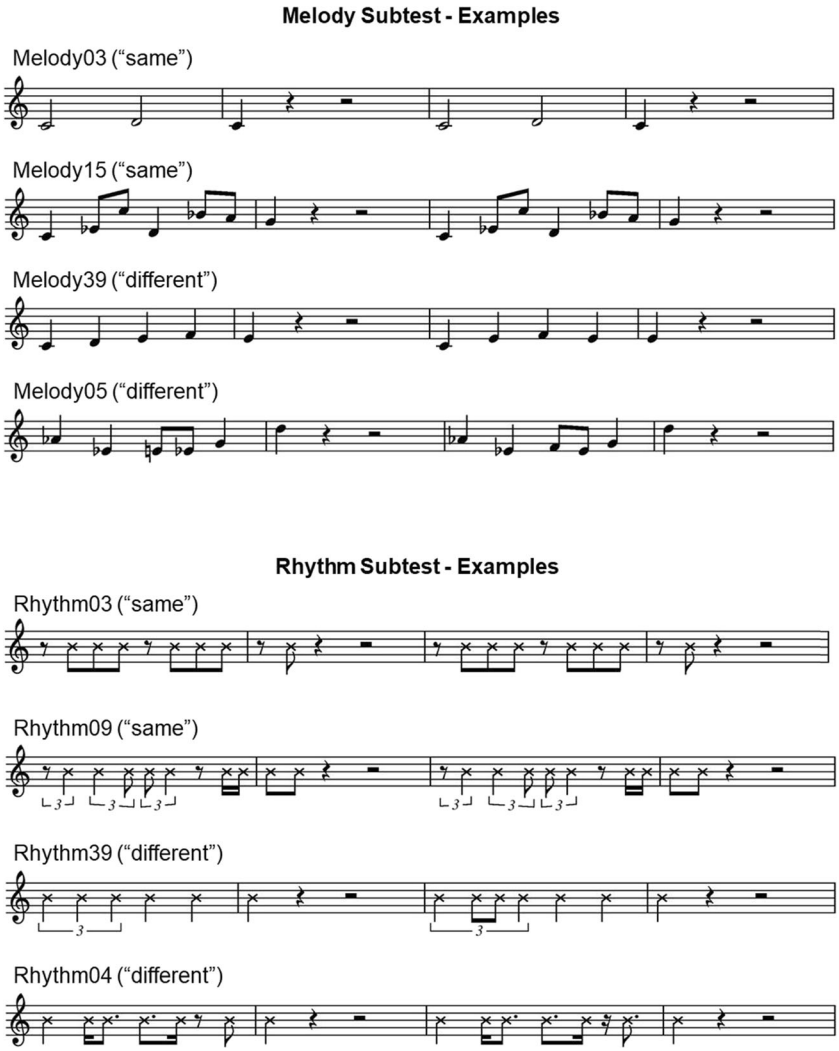

Fig. 1 Example trials from the MET Melody and Rhythm subtests. Reprinted by permission from Springer, *Behavior Research Methods*, "The Musical Ear Test: Norms and correlates from large sample of Canadian undergraduates," Swaminathan, Kragness, & Schellenberg (2021), advance online publication, 11 March 2021, doi: 10.3758/s13428-020-01528-8

select one of seven alternatives from an ordinal scale that varies from item to item. For example, the scale for the statement *I listen attentively to music for …* had options ranging from 1 (*0 - 15 min per day*) to 7 (*4 hours or more per day*).

For European-Portuguese participants, we created an online version of a published translation of the Gold-MSI that has good psychometric properties (Lima et al., 2020). For the Italian translation, items from the original English version were translated to Italian independently by two translators, both of whom were native speakers of Italian, fluent in English, experienced in translating questionnaires, and experts in the psychology of music. The goal was conceptual equivalence rather than a literal translation. Discrepancies between translations were resolved by discussion to create a single version, which was, in turn, evaluated by two independent colleagues for clarity of expression and whether the translation from English was appropriate. The Italian version was then back-translated by a native speaker of English who was fluent in Italian and a scholar of psychology and music. Inconsistencies between the back-translation and the original Gold-MSI were discussed and resolved among the three translators, who also consulted with two additional experts from the discipline. Finally, 10 participants completed the Italian translation of the Gold-MSI and confirmed that the items were clear.

For the Brazilian-Portuguese version, a native speaker, who was also fluent in English and an expert in the psychology of music, made minor modifications to the European-Portuguese version. To ensure that each modification was consistent with the original Gold-MSI, she first checked the English version. Such modifications included the progressive tense (*I am hearing* translated to *estou ouvindo* instead of *estou a ouvir*), the second-person pronoun (replacing *tu* with *você*), some Brazilian-Portuguese idioms, and minor changes in spelling.

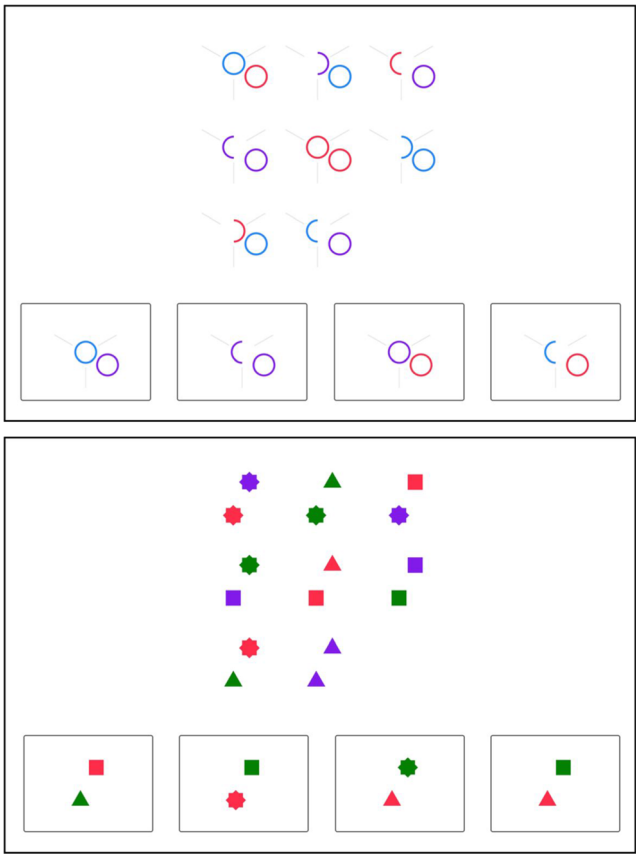

Fig. 2 Two example trials from the Matrix Reasoning Item Bank (MaRs-IB). The third and fourth options are the correct responses for the upper and lower examples, respectively

Cronbach's alphas for the entire sample and for the previously unpublished (Italian and Brazilian-Portuguese) translations of the Gold-MSI are provided in Supplementary Table 1. In general, internal reliability was similar to the comparison sample (Lima et al., 2020), except for a lower alpha in the present sample for the Emotions subtest. Internal reliability was maintained for the previously unpublished translations.

Personality Personality traits were evaluated with the Big Five Inventory (BFI). The BFI is a self-report questionnaire with 44 items that assess five dimensions of personality: openness to experience (10 items), conscientiousness (9 items), extroversion (8 items), agreeableness (9 items), and neuroticism (8 items). Items are mixed in terms of presentation order. Participants rated how much each expression describes them using a five-point rating scale (1 = *disagree strongly*, 5 = *agree strongly*).

The BFI was published initially in English (John & Srivastava, 1999), and subsequently translated into European-Portuguese (Brito-Costa et al., 2015) and Italian (Ubbiali et al., 2013). We created a Brazilian-Portuguese version by modifying the European-Portuguese version, double-checking the original English version for fidelity. Cronbach's alphas for the BFI were acceptable and are provided in Supplementary Table 2.

Mind-wandering As a measure of sustained attention and ability to focus, participants completed the Mind-Wandering Questionnaire (MWQ, Mrazek et al., 2013), a five-item scale with good psychometric properties that evaluates trait levels of mind-wandering (e.g., *I have difficulty maintaining focus on simple or repetitive work*). Participants rated how much they agreed with each sentence on a scale that ranged from 1 (*almost never*) to 6 (*almost always*). Cronbach's alphas for the MWQ were good and are provided in Supplementary Table 2.

Procedure

Participants completed all tasks and questionnaires in one testing session. Access to the experiment was initially provided with a hyperlink posted on social media (e.g., Facebook, Twitter, LinkedIn), which was accompanied by a brief description of the study, including its duration of approximately 40 min. The description also specified that participants should complete the testing session in a quiet room with a stable Internet connection, use headphones, and turn off sound notifications from other devices and applications (e.g., email, phone messages).

The online testing session began with informed consent and some basic demographic questions (e.g., age, gender, education). Participants then completed the self-report questionnaires, which were administered in a fixed order (MWQ, Gold-MSI, and BFI). After the questionnaires, participants were tested on the MaRs-IB and finally the MET. At the end of the study, participants were given feedback about their scores on the personality, musical sophistication, and musical ability measures. A final open-ended question asked participants to describe any problems that might have occurred during the testing session. Some participants reported minor technical difficulties, related primarily to the stability of their Internet connection, but there were otherwise no systematic problems.

Results

The complete data file is provided in the Supplementary Materials. As in the reports from the comparison samples (Lima et al., 2020; Swaminathan et al., 2021), the statistical analyses incorporated standard frequentist null-hypothesis testing, as well as Bayesian analyses conducted with JASP version 0.14.1 (JASP Team, 2020) using default priors.[1] Because of the large sample, very small effects were

[1] Correlations, stretched beta prior width = 1; *t* tests, zero-centered Cauchy prior with scale parameter 0.707; linear regressions, JZS prior of $r = .354$; Wagenmakers et al., 2018a, 2018b; Wagenmakers et al., 2016).

statistically significant with null-hypothesis testing. For example, with $N = 608$, correlations greater than .08 in absolute value were significant with $p < .05$. We considered small associations to be reliable only if they also passed a conventional threshold for what is considered *substantial* evidence using Bayesian statistics (Jarosz & Wiley, 2014; Jeffreys, 1961). Specifically, when the Bayes factor (BF_{10}, reported here with three-digit accuracy) was greater than 3.00, the observed data were at least three times as likely under the alternative as the null hypothesis. Lower values ($1.00 < BF_{10} < 3.00$) indicated that the data provided evidence for the alternative hypothesis that was considered to be weak or anecdotal. If $BF_{10} < 1.00$, the observed data provided evidence that favored the null hypothesis in a reciprocal manner (i.e., substantial evidence when $BF_{10} < .333$). More extreme values provided strong ($BF_{10} > 10.0$ or $< .100$), very strong ($BF_{10} > 30.0$ or $< .033$), and decisive ($BF_{10} > 100.0$ or $< .010$) evidence for either the alternative or null hypothesis, respectively.

Initial analyses documented how the present online sample of participants differed from comparison samples in terms of gender, age, and music training. Detailed statistics are provided in the Supplementary Materials. The present sample had a larger proportion of participants who were men, and the mean age was higher than in Swaminathan et al. (2021) but similar to Lima et al. (2020). Mean levels of music training were higher in the present sample than in both comparison samples.

Swaminathan et al. (2021) did not report personality data, and their sample of undergraduates varied minimally in terms of education. Comparisons with the sample from Lima et al. (2020) revealed that the present sample had lower mean levels of education. For personality (Supplementary Table 3), the two samples differed for each trait, with the present sample scoring higher on openness to experience and neuroticism, but lower on agreeableness, extroversion, and conscientiousness.

The main analyses focused on musical ability, musical experience, and their correlates, including demographics (age, gender, education), cognitive ability, personality, and mind-wandering. Pairwise correlations among potential predictors are provided in Supplementary Table 4. We had no hypotheses about the testing language of the online study, and exploratory analyses confirmed that musical ability did not vary as a function of language when individual differences in age, education, cognitive ability, and openness to experience were held constant. In fact, for the Melody subtest, the Rhythm subtest, and Total scores of the MET, the observed data provided substantial evidence for the null hypothesis (all $BF_{10} < .250$). Testing language was not considered further.

Musical expertise

Because of the large number of musicians in the current sample, mean scores were higher than they were in Lima et al. across subtests and the General Factor, $ps < .001$, all $BF_{10} > 100$ (Supplementary Table 1). As in the comparison sample and elsewhere (Müllensiefen et al., 2014), pairwise correlations among Gold-MSI scores were all positive, and the observed data provided decisive evidence for an association in each instance (Supplementary Table 5). Examination of correlations between Gold-MSI scores and potential predictor variables revealed a relatively small number of instances in which the observed data provided substantial or stronger evidence for an association (Supplementary Table 6).

For demographic variables (age, gender, education), there was decisive evidence of a negative association between age and scores on the Emotions subtest. There was also strong evidence that men had more Music Training than women, and substantial evidence for a male advantage on the General Factor. Cognitive ability had no significant associations with Gold-MSI scores, and the observed data provided substantial (or strong) evidence for the null hypothesis for all subtests. As expected, there was strong evidence for a small, negative association between mind-wandering and the Music Training subtest, but mind-wandering was not associated with any other Gold-MSI score. For personality, openness to experience was associated decisively and positively with all Gold-MSI scores ($rs \geq .4$). The observed data also provided decisive and substantial evidence for positive but small associations between extroversion and Singing Abilities, and between agreeableness and Music Training, respectively ($rs \leq .2$).

Musical ability

Statistics from tests of internal reliability for the online MET are provided in Table 1. Cronbach's alphas were virtually identical to those reported by the test's developers (Wallentin et al., 2010b), and higher than those reported in the comparison sample (Swaminathan et al., 2021). Split-half (odd–even) reliabilities (Spearman-Brown formula) were also

Table 1 Reliability statistics, including Cronbach's alpha and split-half (odd-even) correlations (Spearman-Brown formula), for scores on the MET. For comparison purposes, values from two previous reports are provided

	Melody	Rhythm	Total
Current online sample (N=608)			
Cronbach's alpha	.82	.70	.85
Split-half correlation	.84	.75	.87
Swaminathan et al. (2021, N=523)			
Cronbach's alpha	.73	.62	.78
Split-half correlation	.71	.68	.78
Wallentin et al. (2010b, N=60)			
Cronbach's alpha	.82	.69	.85

considerably higher than those reported by Swaminathan et al. In short, the internal reliability of the MET was not compromised by the online testing format.

Descriptive statistics for the Melody, Rhythm, and Total scores are provided in Table 2. For the entire sample, the observed means were higher than those reported by Swaminathan et al. (2021) for the Melody, Rhythm, and Total scores, as confirmed by independent-samples t tests, $ts(1129) = 5.06$, 5.90, and 6.23, respectively, $ps < .001$, all $BF_{10} > 100$. These findings were not meaningful, however, because of sample differences in musicianship. To rectify this problem, we gave separate consideration to individuals with no music training (see Table 2). For these participants, mean performance did not differ from that reported previously on the Melody subtest, $p = .202$, $BF_{10} = .263$, the Rhythm subtest, $p = .053$, $BF_{10} = .725$, or for Total scores, $p = .064$, $BF_{10} = .625$, although evidence favoring the null hypothesis was substantial only for the Melody subtest. In any event, online-generated scores were comparable to in-person scores when they were expected to be comparable.

As one would expect, Melody and Rhythm scores were positively and decisively correlated, $r = .551$, $N = 608$, $p < .001$, $BF_{10} > 100$, with the magnitude of the association no different from that reported by Swaminathan et al. (2021), $r = .489$, $p = .154$, and Wallentin et al. (2010a), $r = .520$, $p = .754$.[2] As in the earlier reports, the data provided substantial evidence that performance did not differ between subtests, $BF_{10} = .214$.

Demographics, cognitive ability, mind-wandering, and personality

Correlations between MET scores and demographic variables, cognitive ability, mind-wandering, and personality are provided in Table 3. The observed data provided decisive evidence that as listeners increased in age, education, or cognitive ability, performance on the MET (i.e., Melody, Rhythm, and Total scores) tended to improve as well. The one exception was the association between cognitive ability and Melody scores, for which the data provided substantial rather than decisive evidence. The correlation with cognitive ability was also higher for the Rhythm than for the Melody subtest, $z = 2.87$, $p = .004$.

For mind-wandering, there was substantial evidence for a negative association with scores on the Melody subtest, but no evidence of an association with Rhythm or Total scores. Nevertheless, the magnitude of the association was not significantly stronger for Melody than for Rhythm, $p > .1$. For personality, the observed data provided decisive evidence for positive associations between openness to experience and MET performance, but no evidence for associations with any other personality variable. In fact, all Bayes factors were below 1 with a single exception, and for two personality traits (conscientiousness, extroversion), the observed data provided substantial evidence for the null hypothesis.

[2] Comparisons of the magnitude of correlations were conducted with Psychometrica (https://www.psychometrica.de/correlation.html).

Table 2 Descriptive statistics for scores on the MET. Melody and Rhythm scores were calculated from 52 trials. Total scores were calculated from 104 trials. For comparison purposes, values from Swaminathan et al. (2021) are provided

	Current online sample			Swaminathan et al. (2021)		
Whole sample						
	N	*M*	*SD*	*N*	*M*	*SD*
Melody	608	37.88	6.60	523	36.05	5.36
Rhythm	608	38.29	5.35	523	36.47	4.94
Total	608	76.17	10.54	523	72.52	8.89
No music training						
	n	*M*	*SD*	*n*	*M*	*SD*
Melody	151	34.91	6.44	189	34.15	4.41
Rhythm	151	36.66	5.79	189	35.56	4.68
Total	151	71.56	10.90	189	69.71	7.48

Table 3 Pairwise associations (Pearson correlations and Bayes factors) between scores on the MET and demographic variables, cognitive ability, mind-wandering, and personality

		Melody	Rhythm	Total
Age	*r*	.206	.167	.214
	BF_{10}	>100	>100	>100
Gender	*r*	.099	.029	.077
	BF_{10}	1.01	0.066	0.306
Education	*r*	.209	.200	.232
	BF_{10}	>100	>100	>100
Cognitive ability	*r*	.131	.239	.204
	BF_{10}	9.84	>100	>100
Mind-wandering	*r*	-0.122	-0.060	-0.107
	BF_{10}	4.60	0.153	1.63
Openness	*r*	.241	.182	.243
	BF_{10}	>100	>100	>100
Conscientiousness	*r*	.068	.030	.058
	BF_{10}	0.210	0.067	0.142
Extroversion	*r*	.065	.069	.076
	BF_{10}	0.180	0.218	0.288
Agreeableness	*r*	.092	.060	.088
	BF_{10}	0.650	0.151	0.527
Neuroticism	*r*	-0.101	-0.018	-0.072
	BF_{10}	1.11	0.056	0.245

Note. Gender was dummy-coded (female = 0, male = 1). *N*s = 608 except for education, $n = 566$.

Table 4 Pairwise associations (Pearson correlations and Bayes factors) between scores on the MET and scores on the Gold-MSI ($N = 608$)

		Melody	Rhythm	Total
Active Engagement	r	.303	.186	.284
	BF_{10}	>100	>100	>100
Perceptual Abilities	r	.459	.320	.450
	BF_{10}	>100	>100	>100
Music Training	r	.491	.296	.458
	BF_{10}	>100	>100	>100
Singing Abilities	r	.406	.259	.386
	BF_{10}	>100	>100	>100
Emotions	r	.206	.141	.201
	BF_{10}	>100	22.5	>100
General Factor	r	.504	.307	.471
	BF_{10}	>100	>100	>100

Musical expertise and music training

Our main tests of construct validity involved correlations between scores on the MET and those from the subtests and General Factor from the Gold-MSI, which are provided in Table 4. All correlations were positive and statistically significant, with $p < .001$, with the observed data providing decisive evidence for an association in each instance, except for the association between the Emotions subtest and Rhythm scores, which was strong but not decisive.

In the comparison sample (Swaminathan et al., 2021), music training proved to be a better predictor of Melody than of Rhythm scores. Our Gold-MSI scores showed a similar pattern. For Perceptual Abilities, Music Training, Singing Abilities, and the General Factor, correlations with the Melody subtest were higher than those for the Rhythm subtest, zs > 4, ps < .001. The same finding was weaker yet still evident for Active Engagement, $z = 3.16$, $p = .002$, but not for the Emotions subtest, $p = .086$.

Additional analyses focused solely on the Music Training subtest. Associations between Music Training and MET scores (see Table 4) were higher than those in the comparison sample (Swaminathan et al., 2021), which could be due to differences in how training was measured and/or a consequence of greater variability due to the higher proportion of musicians in the present sample. The correlations were somewhat lower than correlations between MET scores and current daily practice reported by Wallentin et al. (2010a, Experiment 3), a likely consequence of differences in measurement.

We also asked whether performance on the MET was associated with the age at which music training began. As in Swaminathan et al. (2021), we considered only participants who had any training ($n = 415$) and divided them into two groups: those who started by age 7—*early starters* ($n = 120$)—and those who started at an older age—*late starters* ($n = 295$). This split was theoretically motivated, based on the proposal of a sensitive period that extends up to 7 years of age, during which plasticity is greater and music training is presumed to have a stronger impact on development (Penhune, 2019, 2020; Penhune & De Villiers-Sidani, 2014).

The results were similar to those reported in the comparison sample (Swaminathan et al., 2021). Early starters had higher scores than late starters on the Melody subtest, $t(413) = 3.18$, $p = .002$, $BF_{10} = 14.7$, and on Total scores, $t(413) = 2.96$, $p = .003$, $BF_{10} = 7.82$, but not on the Rhythm subtest, $p = .076$, $BF_{10} = .543$. Nevertheless, early starters also had more Music Training, $t(413) = 4.11$, $p < .001$, $BF_{10} > 100$. When Music Training was held constant, the advantage for early starters disappeared for the Melody subtest, $p = .078$, $BF_{10} = .577$, and for Total scores, $p = .083$, $BF_{10} = .527$, although the observed data did not provide strong evidence for the null hypothesis.

Multiple regression analysis

In the final set of analyses, we used multiple regression to determine which correlates made independent contributions in predicting performance on the MET. Specifically, we modeled MET Melody, Rhythm, and Total scores from a linear combination of variables, each of which had a reliable simple association with MET scores: age, education, cognitive ability, mind-wandering, openness to experience, and the Gold-MSI subtests. The results are summarized in Table 5. For the Melody subtest, the Rhythm subtest, and Total scores, the overall model was significant, with independent and positive partial associations with age, education, cognitive ability, and the Perceptual Abilities and Music Training subtests from the Gold-MSI.

In the Bayesian counterpart to multiple regression, we first identified which model—out of all possible models—was most likely given the observed data. For the Melody subtest and for Total scores, it was a model that included age, education, cognitive ability, Perceptual Abilities, and Music Training—a finding that corroborated the frequentist results. We calculated a Bayes factor for each predictor by removing them from the model one at a time. As shown in Table 5, the observed data provided decisive evidence for the inclusion of Perceptual Abilities and Music Training in the model, and very strong (Melody) or decisive (Total) evidence for including cognitive ability and age. For education, however, the Bayes factor was less than 3. We calculated BF_{10} for the other (excluded) five variables by adding each to the model one at a time. For each variable, the observed data provided substantial evidence for the null hypothesis. In other words, the observed data were more likely with a model that did *not* include these variables.

Table 5 Multiple regression results predicting MET scores from age, education, openness to experience, cognitive ability, mind-wandering, and the five Gold-MSI subtests

	Melody			Rhythm			Total		
Model									
R^2	.332			.210			.335		
Adjusted R^2	.320			.196			.323		
$F(10, 555)$	27.63			14.76			27.98		
p	<0.001			<0.001			<0.001		
Predictors									
	ß	p	BF_{10}	ß	p	BF_{10}	ß	p	BF_{10}
Age	.154	<0.001	610.7	.159	<0.001	>100	.177	<0.001	>100
Education	.098	.016	1.43	.089	.045	0.760	.107	.009	2.14
Cognitive ability	.145	<0.001	>100	.259	<0.001	>100	.222	<0.001	>100
Mind-wandering	.010	.802	0.129	.013	.751	0.146	.013	.739	0.131
Openness	-0.027	.523	0.129	-0.005	.918	0.160	-0.019	.648	0.125
Active engagement	.049	.347	0.218	.077	.174	0.306	.070	.178	0.357
Perceptual abilities	.177	.003	>100	.174	.008	>100	.199	<0.001	>100
Music training	.305	<0.001	>100	.128	.019	6.62	.256	<0.001	>100
Singing abilities	.053	.337	0.232	-0.019	.749	0.146	.023	.673	0.155
Emotions	-0.002	.972	0.140	-0.006	.908	0.169	-0.004	.931	0.148

For the Rhythm subtest, the best model of the data included age, cognitive ability, Perceptual Abilities, and Music Training. The observed data provided decisive evidence for the inclusion of age, cognitive ability, and Perceptual Abilities in the model, but only substantial evidence for including Music Training. For the other six variables, the observed data provide substantial evidence for the null hypothesis with one exception: they were more or less equally likely with a model that included or excluded education.

Discussion

We sought to determine whether an established and validated test of musical ability could be administered successfully online. Although approximately 20% of the sample who started the testing session did not complete it or provide usable data, this level of attrition is not surprising, because there was no compensation or incentive for participants to complete the session, other than to receive feedback about their personality, musical expertise, and musical ability. Moreover, the testing session was relatively long and, unlike in a laboratory, there were no research assistants to witness a participant's decision to discontinue. In any event, the findings were otherwise unequivocally positive. Indeed, the results for the MET were both novel and noteworthy because it is an objective listening test of musical ability that, to our knowledge, has not been adapted previously for online testing.

The Gold-MSI served as our main variable for testing construct validity and as a proof of concept—that the present sample of online participants would respond similarly to a sample of participants tested in a more traditional format (Lima et al., 2020). Indeed, response patterns to the online Gold-MSI were very similar to those reported previously. For example, the internal reliability of the test was similar across formats except for the Emotions subtest. As in the earlier study, age correlated negatively with the Emotions subtest, although Lima et al. found a negative correlation between age and *all* Gold-MSI subtests. Discrepancies in response patterns between samples could stem from differences in music training. Compared to the previous study, we had a larger subsample of participants with very high levels of music education; one-quarter of our sample (25.6%) had 10 or more years of music lessons, whereas in Lima et al., the figure was closer to one-twentieth (5.6%). Because increases in musical experience must be accompanied by increases in age, a negative association between age and Gold-MSI scores would be less likely in our online sample. Despite these differences in samples, correlations among Gold-MSI subtests, and between Gold-MSI scores and personality variables, were similar across testing formats.

One null finding was that there was little evidence of an association between cognitive ability and the Music Training subtest from the Gold-MSI. In childhood, music training is often correlated positively with cognitive ability (Corrigall et al., 2013; Corrigall & Schellenberg, 2015; Kragness et al.,

2021; Schellenberg, 2006, 2011; Schellenberg & Mankarious, 2012; Swaminathan & Schellenberg, 2020). In adulthood, however, such associations tend to be weaker (Lima & Castro, 2011; Schellenberg, 2006). When matrix-type tests of cognitive ability, such as Raven's test and the test used in the present sample (MaRs-IB), are given to students from an introductory psychology course, positive associations with music training are evident in some instances (Swaminathan et al., 2017, 2018, 2021; Swaminathan & Schellenberg, 2018) but not in others (Schellenberg & Moreno, 2010; Swaminathan & Schellenberg, 2017). These associations may become less likely in samples of older participants with a large proportion of professional musicians (Lima & Castro, 2011).

Turning now to our main focus, the MET, the internal reliability of the online version proved to be similar to, perhaps even better than, in-person administration (Wallentin et al., 2010b; Swaminathan et al., 2021). Other results confirmed that (1) the correlation between Melody and Rhythm subtests did not differ across formats, (2) there was no difference in performance between subtests, and (3) when the present and comparison samples were equated for music training by focusing solely on participants with no training, average levels of performance were similar. Moreover, as in the comparison sample, there were no gender differences in performance on the MET. Finally, as in other samples, performance was strongly associated with openness to experience, but not with other dimensions of personality (Greenberg et al., 2015; McCrae & Greenberg, 2014; Swaminathan & Schellenberg, 2018; Thomas et al., 2016). In short, online testing did not compromise the reliability and validity of the MET.

Strong evidence of construct validity for our online version of the MET came from positive associations with scores on the Gold-MSI. Previous in-person studies documented that as the degree of musicianship and amount of practice (Wallentin et al., 2010a) or duration of music training (Swaminathan et al., 2021) increases, so does performance on the MET. In the present investigation, associations with Music Training as measured by the Gold-MSI were somewhat higher than those of the comparison sample (Swaminathan et al., 2021), which we attribute to the relatively high variability in music training and the high proportion of professional musicians tested online. We also found positive associations between MET scores and other aspects of self-reported musical expertise measured by the Gold-MSI, namely Active Engagement, Emotions, Perceptual Abilities, and Singing Abilities. In the Gold-MSI validation study, Müllensiefen et al. (2014) reported a comparable pattern of associations using short beat alignment and melodic memory tasks. Our results extended these associations, indicating that musical skills and experience are multifaceted, and not limited to music lessons or playing an instrument. Moreover, even though the musical skills tested by the MET are based on auditory short-term (working) memory and perceptual discrimination, performance was predictive of a broad range of musical behaviors and expertise.

As in the comparison sample, we found no association between musical abilities and age of onset of music lessons after duration of music training was held constant. This finding raises the possibility that proposals of plasticity effects arising from early music training (Penhune, 2019, 2020; Penhune & De Villiers-Sidani, 2014) may be exaggerated. Indeed, longitudinal evidence in childhood shows that musical ability is independent of music training when levels of musical ability measured 5 years previously are taken into account (Kragness et al., 2021). Nevertheless, other findings reveal behavioral advantages and structural brain differences as a consequence of early training, even after accounting for duration of training (Bailey et al., 2014; Bailey & Penhune, 2010, 2012, 2013). Perhaps early onset of music training explains some musical abilities, such as rhythm synchronization and production abilities, but not other abilities, such as those measured by the MET.

As noted, one advantage of online recruitment is that it allowed for a large sample of motivated individuals, including many who likely participated because they identified as working musicians or musician-academics. Our sample was also heterogeneous in terms of age and education, which tend to vary minimally when participants are recruited from undergraduate courses in introductory psychology, as in the MET comparison sample (Swaminathan et al., 2021). Substantial variance in education meant that we had two variables to represent cognitive ability: the objective test as well as self-reports of education. The status of age and its relation to cognition is more ambiguous, because some abilities, such as processing speed, start to decline relatively early in life, whereas others continue to peak until after age 40 (Hartshorne & Germine, 2015). In any event, age, education and our online measure of cognitive ability were predictive of performance on the MET. In the comparison sample, MET scores correlated positively with three different measures of cognitive ability: digit span forward, digit span backward, and Raven's tests. Thus, as with virtually any specific cognitive ability, individual differences in musical ability vary positively with general ability (Carroll, 1993), whether they are measured in person or online.

Although the association between MET scores and cognitive abilities was consistent with previous research (e.g., Swaminathan et al., 2017, 2018, 2021; Swaminathan & Schellenberg, 2018), and strong even when other variables were held constant (Table 5), cognitive ability was a better predictor of scores on the Rhythm compared to the Melody subtest. Swaminathan et al. (2021, Table 8) also found evidence that general ability (i.e., working memory as measured by digit span backward) was a better predictor of Rhythm than of Melody scores. By contrast, music training was a better predictor of Melody compared to Rhythm in the online *and*

in-person samples, and this difference extended to other aspects of musical expertise measured by the Gold-MSI, specifically Active Engagement, Perceptual Abilities, Singing Ability, and the General Factor. In other words, performance on the Melody subtest appears to rely more on individual differences in exposure to music, whereas performance on the Rhythm subtest is more strongly associated with nonmusical individual differences. Swaminathan et al. (2021) suggested that this result might stem from the fact that the Rhythm subtest taps into a universal feature of music, whereas performance on the Melody subtest is more strongly influenced by exposure to pitch structures that are specific to Western music. Even in early childhood, 1 year of intensive music training improves melody discrimination more than it improves rhythm discrimination (Ilari et al., 2016).

Performance on the Melody subtest but not the Rhythm subtest was also linked to a lower level of mind-wandering, although this association disappeared when other predictors of Melody scores were held constant. In one previous study (Wang et al., 2015), highly trained musicians had an enhanced ability to sustain attention during a temporal discrimination task (but not in a visual discrimination task), and this advantage remained evident when cognitive ability was held constant. The association between musical ability and mind-wandering or sustained attention could be examined in more detail in future research.

Because the Gold-MSI subscales had considerable overlap (Supplementary Table 5), the multiple regression analyses served to identify which subscales made independent contributions to predicting performance on the MET. In addition to the Music Training subscale, the Perceptual Abilities subscale was a robust predictor of Melody, Rhythm, and Total scores, and, in the case of Rhythm, even superior to Music Training. This finding is indicative of participants' meta-cognitive awareness of their musical ability: Individual differences in self-reports of music perception skills, measured before taking the MET, correlated with musical abilities measured subsequently and objectively.

The present study also had limitations. Although we asked participants to perform the experiment in a quiet environment and to avoid distractions, Internet testing made it difficult to control for extraneous sounds or potential interruptions, which remain a major challenge for online testing in general, and for auditory research in particular. Moreover, we did not include a task to ensure that participants used headphones (Milne et al., 2020; Woods et al., 2017). Although we strongly recommended that they use them throughout the experiment, it was not possible to verify whether they did.

In sum, the online version of the MET showed good internal reliability and appropriate levels of performance. Strong associations between the accuracy on the MET and musical sophistication and training, especially for the Melody subtest, were also consistent with studies using in-person testing of MET (Swaminathan et al., 2021). Finally, as expected, scores from online administration correlated with personality (openness to experience), cognitive ability, and mind-wandering. Online testing also had advantages compared to the traditional in-lab testing, which have been noted by others (e.g., Casler et al., 2013; Gosling et al., 2004). For example, online recruitment allowed us to obtain a larger and more diverse sample compared to previous studies on musical abilities, including participants from different nationalities, a large number of professional musicians as well as nonmusicians, and participants who varied widely in age. Finally, the online format made it possible to recruit participants and collect data in a very short time (approximately 1 month), because we were not limited by the space and time constraints of the laboratory.

To conclude, our findings showed that online administration of MET is a valid and reliable alternative to traditional in-person measurement of musical abilities. With greater worldwide access to the Internet, and in-person restrictions imposed by the COVID-19 pandemic, there has been a growing interest in the development of Internet methods. This study contributes to the growing literature on the utility of online testing as an alternative, or complement, to laboratory testing for psychological research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01641-2.

Funding Funded by the Fundação para a Ciência e a Tecnologia (FCT; grant PTDC/PSI-GER/28274/2017 awarded to C.F.L., and a Scientific Employment Stimulus grant to E.G.S). **Open Practices Statement** The data are included as an electronic file in the Supplementary Materials. The study was not pre-registered.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Asztalos, K., & Csapó, B. (2014). Online assessment of musical abilities in Hungarian primary schools—results of first, third and fifth grade students. *Bulletin of the International Kodály Society, 39*(1), 3–14.

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. A., & 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, Article 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*. October House.

Birnbaum, M. H. (2004). Methodological and ethical issues in conducting social psychological research via the Internet. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *Handbook of methods in social psychology* (pp. 359–382). Sage Publications, Inc.

Brito-Costa, S., Bem-Haja, P., Moisão, A., Alberty, A., Castro, F. V., & De Almeida, H. (2015). Psychometric properties of Portuguese version of Big Five Inventory (BFI). *International Journal of Developmental and Educational Psychology* [*INFAD Revista de Psicologia*], *1*(2), 83–94. https://doi.org/10.17060/ijodaep.2015.n2.v1.325

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 University Press. https://doi.org/10.1017/CBO9780511571312

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Psychology*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Chetverikov, A., & Upravitelev, P. (2015). Online versus offline: The Web as a medium for response time data collection. *Behavior Research Methods*, *48*(3), 1086–1099. https://doi.org/10.3758/s13428-015-0632-x

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, *6*(10), Article 190232. https://doi.org/10.1098/rsos.190232

Cooper, P. K. (2019). It's all in your head: A meta-analysis on the effects of music training on cognitive measures in schoolchildren. *International Journal of Music Education*, *38*(3), 321–336. https://doi.org/10.1177/0255761419881495

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

Corrigall, K. A., & Schellenberg, E. G. (2015). Predicting who takes music lessons: Parent and child characteristics. *Frontiers in Psychology*, *6*, Article 282. https://doi.org/10.3389/fpsyg.2015.00282

Corrigall, K. A., Schellenberg, E. G., & Misura, N. A. (2013). Music training, cognition, and personality. *Frontiers in Psychology*, *4*, Article 222. https://doi.org/10.3389/fpsyg.2013.00222

Dandurand, F., Bowen, M., & Shultz, T. R. (2004). Learning by imitation, reinforcement and verbal rules in problem-solving tasks. In J. Triesch & T. Jebara (Eds.). *Proceedings of the 2004 International Conference on Development and Learning* (pp. 88–95). UCSD Institute for Neural Computation.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. https://doi.org/10.3758/brm.40.2.428

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners. *Psychological Science*, *27*(6), 848–858. https://doi.org/10.1177/0956797616639301

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*. Houghton Mifflin.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*(1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*(2), 93–104. https://doi.org/10.1037/0003-066x.59.2.93

Greenberg, D. M., Müllensiefen, D., Lamb, M. E., & Rentfrow, P. J. (2015). Personality predicts musical sophistication. *Journal of Research in Personality, 58*, 154–158. https://doi.org/10.1016/j.jrp.2015.06.002

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, *52*(6), 2283–2286. https://doi.org/10.3758/s13428-020-01395-3

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological science*, *26*(4), 433–443. https://doi.org/10.1177/0956797614567339

Houben, K., & Wiers, R. W. (2008). Measuring implicit alcohol associations via the Internet: Validation of Web-based implicit association tests. *Behavior Research Methods*, *40*(4), 1134–1143. https://doi.org/10.3758/brm.40.4.1134

Ilari, B. S., Keller, P., Damasio, H., & Habibi, A. (2016). The development of musical skills of underprivileged children over the course of 1 year: A study in the context of an El Sistema-inspired program. *Frontiers in Psychology, 7*:62. https://doi.org/10.3389/fpsyg.2016.00062

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. (2020). *JASP* (Version 0.14.1) [Computer software]. https://jasp-stats.org/

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

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Kragness, H. E., Swaminathan, S., Cirelli, L. K., & Schellenberg, E. G. (2021). Individual differences in musical ability are stable over time in childhood. *Developmental Science*. Advance online publication. https://doi.org/10.1111/desc.13081

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). Academic Press.

Kuckelkorn, K., de Manzano, Ö., & Ullén, F. (2021). Musical expertise and personality—differences related to occupational choice and instrument categories. *Personality and Individual Differences*, *173*, Article 110573. https://doi.org/10.1016/j.paid.2020.110573

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), Article e52508. https://doi.org/10.1371/journal.pone.0052508

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*(5), 1021–1031. https://doi.org/10.1037/a0024521

Lima, C. F., Correia, A. I., Müllensiefen, D., & Castro, S. L. (2020). Goldsmiths Musical Sophistication Index (Gold-MSI): Portuguese

version and associations with socio-demographic factors, personality and music preferences. *Psychology of Music*, *48*(3), 376–388. https://doi.org/10.1177/0305735618801997

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., & Greenberg, D. M. (2014). Openness to experience. In D. K. Simonton (Ed.), *Handbook of genius* (pp. 222–243). Wiley-Blackwell

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-020-01514-0

Moreno, S., Bialystok, E., Barac, R., Schellenberg, E. G., Cepeda, N. J., & Chau, T. (2011). Short-term music training enhances verbal intelligence and executive function. *Psychological Science*, *22*(11), 1425–1433. https://doi.org/10.1177/0956797611416999

Mrazek, M. D., Phillips, D. T., Franklin, M. S., Broadway, J. M., & Schooler, J. W. (2013). Young and restless: validation of the Mind-Wandering Questionnaire (MWQ) reveals disruptive impact of mind-wandering for youth. *Frontiers in Psychology*, *4*, Article 560 https://doi.org/10.3389/fpsyg.2013.00560

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), Article e89642. https://doi.org/10.1371/journal.pone.0089642

Nussenbaum, K., Scheuplein, M., Phaneuf, C. M., Evans, M. D., & Hartley, C. A. (2020). Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology*, *6*(1), Article 17213. https://doi.org/10.1525/collabra.17213

Penhune, V. B. (2019). Musical expertise and brain structure: The causes and consequences of training. In M. H. Thaut & D. A. Hedges, (Eds.), *The Oxford handbook of music and the brain* (pp. 417–438). 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. B., & de Villers-Sidani, E. (2014). Time for new thinking about sensitive periods. *Frontiers in Systems Neuroscience*, *8*, Article 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

Peretz, I., Gosselin, N., Nan, Y., Caron-Caplette, E., Trehub, S. E., & Béland, R. (2013). A novel tool for evaluating children's musical abilities across age and culture. *Frontiers in Systems Neuroscience, 7*:30. https://doi.org/10.3389/fnsys.2013.00030

Raven, J. C. (1965). *Advanced Progressive Matrices,* . Psychological Corporation.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 496–505. https://doi.org/10.3758/bf03210601

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. (2011). Music lessons, emotional intelligence, and IQ. *Music Perception*, *29*(2), 185–194. https://doi.org/10.1525/mp.2011.29.2.185

Schellenberg, E. G., & Mankarious, M. (2012). Music training and emotion comprehension in childhood. *Emotion*, *12*(5), 887–891. https://doi.org/10.1037/a0027971

Schellenberg, E. G., & Moreno, S. (2010). Music lessons, pitch processing, and g. *Psychology of Music*, *38*(2), 209–221. https://doi.org/10.1177/0305735609339473

Schaal, N. K., Banissy, M. J., & Lange, K. (2015). The rhythm span task: Comparing memory capacity for musical rhythms in musicians and non-musicians. *Journal of New Music Research, 44*(1), 3–10. https://doi.org/10.1080/09298215.2014.937724

Seashore, C. (1919). *The psychology of musical talent*. Holt.

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

Swaminathan, S., Kragness, H. E., & Schellenberg, E. G. (2021). The Musical Ear Test: Norms and correlates from a large sample of Canadian undergraduates. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-020-01528-8

Swaminathan, S., & Schellenberg, E. G. (2020). Musical ability, music training, and language ability in childhood. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2340–2348. https://doi.org/10.1037/xlm0000798

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*(1), Article 9223. https://doi.org/10.1038/s41598-018-27571-2

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

Taherbhai, H., Seo, D., & Bowman, T. (2012). Comparison of paper–pencil and online performances of students with learning disabilities. *British Educational Research Journal*, *38*(1), 61–74. https://doi.org/10.1080/01411926.2010.526193

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

Ubbiali, A., Chiorri, C., Hampton, P., & Donati, D. (2013). Italian Big Five Inventory. Psychometric properties of the Italian adaptation of the Big Five Inventory (BFI). *Bollettino di Psicologia Applicata, 266*(59), 37–48.

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

Wagenmakers, E., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2018a). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey,

R. D. (2018b). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods, 48*(2), 413–426. https://doi.org/10.3758/s13428-015-0593-0

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) (2010) 188–196]. *Learning and Individual Differences, 20*(6), 705. https://doi.org/10.1016/j.lindif.2010.10.001

Wang, X., Ossher, L., & Reuter-Lorenz, P. A. (2015). Examining the relationship between skilled music training and attention. *Consciousness and Cognition*, *36*, 169–179. https://doi.org/10.1016/j.concog.2015.06.014

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

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, *1400*(1), 33–45. https://doi.org/10.1111/nyas.13410