The Japanese translation of the Gold-MSI: Adaptation and validation of the self-report questionnaire of musical sophistication

Makiko Sadakata iD
University of Amsterdam, The Netherlands

Yasumasa Yamaguchi
Tohoku University, Japan

Chie Ohsawa
Mukogawa Women's University, Japan

Masaki Matsubara
University of Tsukuba, Japan

Hiroko Terasawa
University of Tsukuba, Japan

Andres von Schnehen iD
Université de Lille, France

Daniel Müllensiefen
University of London, UK

Kaoru Sekiyama
Kyoto University, Japan

Abstract

This study presents a Japanese translation of the Goldsmiths Musical Sophistication Index (Gold-MSI). The index consists of 38 self-report questions and provides a general sophistication score as well as subscale

Corresponding author:
Makiko Sadakata, Music Department, Institute for Logic, Language and Computation, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands.
Email: m.sadakata@uva.nl

scores for Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions. The validation of the translation with 689 native Japanese speakers indicated excellent internal consistency and test–retest reliability. Confirmatory factor analysis revealed that the bifactor model structure formulated by the original study of Gold-MSI is maintained reasonably in our data. The strengths of the Gold-MSI self-report inventory are (1) it offers a multifaceted view of musical sophistication, (2) a subset of five subscales can be used to measure different aspects of musical sophistication independently, and (3) the ease of administration as it is a self-report questionnaire. In view of the fact that this inventory and its translations increasingly contribute to research on musical expertise, skills, and abilities, having a Japanese translation may enhance future research in these areas even further.

Keywords
musicality, musical abilities, test adaptation, reliability

An important predictor variable in much of music research is an individual's musicality, a complex concept that is associated with various aspects of musical abilities, such as musical skills, aptitude, and musicianship. Traditionally, researchers have adopted various criteria to approximate this complex construct. For example, many studies determined individuals' musicianship by using the extent of their musical training or asking if they had been admitted to a music conservatory (e.g., Kolinsky et al., 2009). Although these estimates have been helpful, their thresholds and criteria vary considerably among laboratories, making it challenging to compare data on musical expertise across studies. For example, while many studies use 10 years as a cut-off point of musicianship (e.g., Nazari et al., 2018; Rammsayer & Altenmüller, 2006; Wolpert, 2000), many others may consider their participants musicians if they have 4–5 years of experience (e.g., Aksentijevic et al., 2014; Jimenez & Kuusi, 2018; Panagiotidi & Samartzi, 2013; Sadakata & Sekiyama, 2011; Schiavio & Timmers, 2016; Weijkamp & Sadakata, 2017; Yu et al., 2016) and, notably, some studies included 4, 5, and 6 years of experience as an upper cut-off point for defining non-musicians (Collier & Collier, 2007; Hansen et al., 2013; Madison & Merker, 2005; Olsen et al., 2018; Radvansky et al., 1995; Repp et al., 2012; Wolpert, 2000). Therefore, depending on the study, the same person with about 4–5 years of musical experience could be classified as a musician or non-musician. Besides, conceptualizing musicality with reference to the number of years of training fails to capture the full scope of musical experiences in which people engage. An increasingly common way to assess musicality is to use batteries of musical listening tests, such as the Profile of Music Perception Skills (Law & Zentner, 2012), the Musical Ear Test (Wallentin et al., 2010), and the Swedish Musical Discrimination test (Ullén et al., 2014), to name but a few. Listening tests commonly address different musical skills such as the perception of pitch differences, pitch contours, melody, chord progression, and rhythm. While listening tests provide more objective measures of musical abilities, the number of specific skills they can measure is usually limited and their administration might require specific material or a local or online testing platform. In contrast, self-report questionnaires can cover a wider range of different aspects of expertise (e.g., not only listening abilities) and are easier to administer.

Although existing batteries consider various dimensions of musical perceptual and performance skills, the question remains whether these measures genuinely reflect the individual's musical aptitude or musicality. Musicality is a concept whose multifaceted nature (cf. Hallam, 2010, 2016) makes it difficult to offer a simple definition. Gembris (1997) describes three different phases in the history of research on musical ability. The initial investigation of musical

ability started in the 1800s, mainly focusing on phenomenological aspects of musical listening and performance (e.g., Michaelis et al. as cited in Gembris, 1997). In a second phase during the 20th century, research focused on the development of psychometric measures of musical abilities in the form of test batteries for musical aptitude or talent (e.g., Galton, 1883; Gordon, 1967; Seashore et al., 1960; Wing, 1962). A more recent approach in the late 1980s focused on the cognitive understanding of musical meaning (Sloboda, 1993; Swanwick & Tillman, 1986). However, the discussion of what constitutes musicality continues to the present day. How accurately one can recognize a melody or a rhythm is undoubtedly a part of this concept, but recent views are increasingly incorporating a broader range of skills. For example, Chin and Rickard (2010) advocate taking into account musical engagement aside from specialized skills in performing and perceiving music, which they define as involving "advanced or frequent listening experiences, a commitment to music activities, or a high level of participation in a particular type of musical activity (e.g., social or emotional)" (p. 198). In accordance with such suggestions, more measurement instruments have started to include assessments of various aspects of individual differences in musicality. The Goldsmiths Musical Sophistication Index (Gold-MSI) is one of the most popular (Müllensiefen et al., 2014). Müllensiefen and colleagues coined the term *musical sophistication*, a more neutral and inclusive term that sheds light on observable behaviors, compared to other terms that relate more to an individual's musical potential, such as talent and aptitude. The index consists of 38 self-report questions grouped into the following five subscales: Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions. This set of questions was refined from 153 statements that the authors had extracted from existing studies through a rigorous selection and evaluation procedure. The index was validated by a large number of mainly English-speaking participants and it has demonstrated good internal reliability for each of the five subscales and the general sophistication index, high test–retest reliability, and reliable correlation with a wide range of objective listening ability tests (e.g., Gelding et al., 2021; Harrison, Collins, & Müllensiefen, 2017; Harrison & Müllensiefen, 2018; Larrouy et al., 2019; Lee & Müllensiefen, 2020; MacGregor & Müllensiefen, 2019; Müllensiefen et al., 2014)

The Gold-MSI has been widely used since it was introduced in 2014 and has been translated into several languages, such as Traditional Chinese (Lin et al., 2021), Simplified Chinese (Li et al., 2021), Portuguese (Lima et al., 2020), German (Schaal et al., 2014), and French (Degrave & Dedonder, 2019). Evaluations of these translations show high internal consistency (Degrave & Dedonder, 2019; Lima et al., 2020; Lin et al., 2021; Schaal et al., 2014) and high test–retest reliability (Lima et al., 2020; Lin et al., 2021). In general, the validation data collected by various studies show a satisfactory-to-good fit with the bifactor model structure suggested by the original Gold-MSI (Degrave & Dedonder, 2019; Lima et al., 2020; Lin et al., 2021; Schaal et al., 2014). These demonstrate that the structure and the set of questions put forward to capture musical sophistication by the original Gold-MSI hold for other cultures and languages as well.

There are some batteries developed for testing individuals' musical abilities (i.e., perceptual and performance skills and interpretation of musical pieces) in Japanese, such as Onkenshiki Youji no Ongaku Tekisei Tesuto (Onken Musical Aptitude Test for Young Children, Onken style, Ongaku Shinri Kenkyu Sho, 1969), Takenshiki Ongaku Soshitsu Tesuto (Taken Musical Aptitude Test, Motegi et al., 1959), and New Musical Aptitude Test (Ogawa, Murao, & Mang, 2008). It is interesting to note that many of these batteries are designed to measure the musical abilities of children. Although not many non-Japanese standardized measures have been translated into Japanese so far, a few are available, such as the Bentley Measure of Musical Abilities (Furuichi & Umemoto, 1975), a battery measuring children's musical abilities. For measuring music-related traits in adults, the Profile of Music Perception Skills (Law & Zentner, 2012) and

the Harvard Beat Assessment Test (focusing more on rhythmic aspects, Fuji & Schlaug, 2013) are both available with Japanese instructions. Furthermore, it is also common that individual studies (e.g., Satoh et al., 2011; Zhang et al., 2020) adopt existing perceptual tests with specially translated Japanese instructions, such as the Montreal Battery of Evaluation of Amusia (Peretz et al., 2003), the Musical Aptitude Profile (Gordon, 1967, 1995), the Musical Ear Test (Wallentin et al., 2010), and the Seashore Measure of Musical Talents (Seashore et al., 1960).

Given the current situation, having a Japanese version of the Gold-MSI would facilitate future work within and across research communities. For self-report questionnaires such as the Gold-MSI, a validated translation of instructions and question items is crucial. Thus, the goal of the current study was to translate and validate the Japanese version of the Gold-MSI, hereafter Gold-MSI-J. We report the evaluation of internal reliability, test–retest reliability, inter-factor correlations, confirmatory factor analyses for model comparisons, as well as correlations between the Gold-MSI-J and important demographic variables. We discuss similarities and differences between the validation results of our version of the Gold-MSI and the original English version, as well as other translated versions.

Translation

Two native Japanese speakers fluent in English and familiar with neuropsychological questionnaires created two independent Japanese translations of the Gold-MSI. The two translators then combined both translated questionnaires into one version based on their individual translations. Another two independent Japanese native speakers checked the translations and adjusted their fluency and readability. Finally, one bilingual speaker of English and Japanese translated the Japanese questionnaire back into English again. Differences between the back translation and the original Gold-MSI were resolved unanimously. Whenever necessary, translations were adjusted in such a way as to aim to capture conceptual meaning rather than the literal expression of the question. The Gold-MSI-J can be downloaded from https://shiny.gold-msi.org/.

Methods

Participants

A total of 718 participants took part in the online survey. Among them, we discarded the data of 14 participants because they did not want to be included in the study, 8 because they filled in the questionnaire more than once (excluding the invited retest subgroup), 5 because Japanese was not their native language, and 2 because they gave significantly inconsistent answers to the same questions phrased differently to catch cursory responses (difference greater than 4 scale points).[1] This resulted in a final sample of 689 participants (female = 282, male = 396, other = 5, undisclosed gender = 6). While many of our participants were under 29 years old, we tried to recruit other age groups as much as possible to increase diversity and create an appropriate sample. Instead of the exact age, participants gave categorical responses (the coding of the age group "under 20" as 1, "20–29" as 2, "30–39" as 3, etc.) (mean category age $M = 2.58$). The age distribution and the distribution of the expected highest educational qualifications are illustrated in Figure 1. Because educational qualifications were given as optional questions, not all participants responded to these (see Tables 4 and 5 for the sample size). A question regarding socioeconomic status was asked to a subset of participants ($n = 148$, see Table 5 for details).
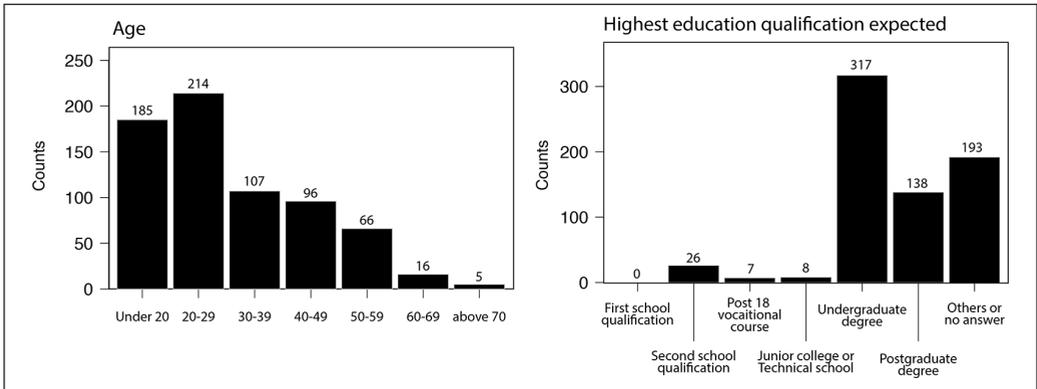
Figure 1. Distribution of age and expected highest educational qualifications.

Retest participants

From among the 689 participants, 23 took the survey again, approximately 3 weeks after their first response (mean duration $M = 21.1$ days, $SD = 1.75$). All retest participants were university students (undergraduate degree) and aged between 20 and 29 years old at the time of testing.

Data collection

Data collection took place during different periods from October 2018 to May 2021. Participants were recruited through institutional and social networks. Multiple web survey platforms were used to collect the data (Google Forms $n = 85$, Qualtrics $n = 456$, CrowdWorks $n = 148$). A total of 150 participants who took part in the survey via CrowdWorks received a small monetary compensation for their contributions (220 yen = approximately 1.7 euros). Due to a technical error, 239 participants answered the question "Ability to Accompany Novel Tune" twice instead of "Judge Others' Singing Ability." The answers to the missing question were estimated by two methods: Full-Information-Maximum-Likelihood (FIML) estimation provided by *lavaan* package (v0.6-9; Rossel, 2012) for fitting the models, and multiple imputation (10 imputations using all other variables) for computing the other statistics that are reported in Tables 1, 2, and 4 using IBM SPSS (v.28.0.1.0). The research was approved by the ethics committee of the Faculty of Humanities of the University of Amsterdam.

Results

Table 1 shows the summary statistics and reliability indicators of the Gold-MSI-J subscales and general musical sophistication factor. The Shapiro–Wilk test confirmed that the scores for each subscale and the general musical sophistication factor were normally distributed. In general, measures of internal consistency and test–retest reliability were high, indicating a strong internal consistency. The test–retest correlation for the subscale Emotions was somewhat lower but still within the acceptable range, which mirrors the results of existing versions of Gold-MSI (e.g., Lin et al., 2021). The inter-factor correlations are shown in Table 2. The order of low to high correlation coefficients in Table 2 is comparable to that of German (Schaal et al., 2014) and French (Degrave & Dedonder, 2019) translations of the Gold-MSI.

Table 1. Summary statistics and indicators of reliable analyses for the Japanese version of the Gold-MSI subscales and general Musical Sophistication scale.

	Mean score (SD)	Score range (Theoretical max.)	Sk.	Kur.	McDonald's ω	Cronbach's α	Gutmann's λ6	r_{tt}
F1 Active Engagement (9 items)	32.39 (9.02)	9–60 (63)	.041	−.227	.817	.811	.806	.92
F2 Perceptual Abilities (9 items)	41.57 (9.62)	9–63 (63)	−.356	−.043	.865	.862	.863	.9
F3 Musical Training (7 items)	19.83 (9.73)	7–49 (49)	.608	−.350	.833	.832	.825	.96
F4 Singing Abilities (7 items)	25.55 (8.31)	7–47 (49)	.048	−.448	.817	.811	.809	.94
F5 Emotions (6 items)	30.52 (5.81)	10–42 (42)	−.568	.434	.746	.742	.729	.73
FG General Musical Sophistication (18 items)	65.81 (19.16)	18–120 (126)	.096	−.293	.899	.898	.915	.95

Note: SD: standard deviation; theoretical max.: theoretical scale maximum; Sk.: skewness; Kur.: kurtosis, r_{tt}: test–retest correlation; FG: General Musical Sophistication.

Table 2. Inter-factor correlations between the five subscales of the Gold-MSI (Pearson).

	Active Engagement	Perceptual Abilities	Musical Training	Singing Abilities
F1 Active Engagement	–	–	–	–
F2 Perceptual Abilities	.508	–	–	–
F3 Musical Training	.451	.617	–	–
F4 Singing Abilities	.486	.743	.616	
F5 Emotions	.664	.534	.369	.466

Note: Reported correlations are all highly significant (Bonferroni corrected $p < .001$).

Subsequently, we investigated whether the bifactor model accounts best for the Japanese questionnaire data, just as in the original study (Müllensiefen et al., 2014) and the Traditional Chinese translation of the Gold-MSI (Lin et al., 2021). Here, four models are compared using confirmatory factor analyses (*lavaan* package, v0.6-9; Rossel, 2012). Model 1 is a hierarchical model that assumes an influence of the general factor on the five factors as well as the influence of these five factors on individual items associated with each. Model 2 is a bifactor model that assumes that each item is influenced by both the general factor and one of the five factors. Models 3 and 4 do not assume a general factor but define five factors, each associated with individual items. While Model 3 does not allow factors to correlate, Model 4 allows for factor intercorrelations. Table 3 summarizes the model fit of each model.

Table 3. Summary results of confirmatory factor analyses comparing four different structural models.

	Model 1	Model 2	Model 3	Model 4
	(Hierarchical)	(Bifactor)	(Simple factors)	(Correlated factors)
χ^2	3,193.24	2,565.11	4,678.662	3,027.15
df	660	627	665	655
AIC	90,739.29	58,281.08	92,214.710	90,583.20
BIC	91,278.98	58,949.53	92,731.727	91,145.57
CFI	.768	.823	.633	.783
TLI	.753	.801	.612	.767
RMSEA	.075	.067	.094	.073
SRMR	.075	.066	.236	.071

Note: AIC: Akaike information criterion, BIC: Bayesian information criterion, CFI: Bentler's comparative fit index, TLI: Tucker–Lewis index, RMSEA: root mean square error of approximation, SRMR: standardized root mean square residual. R:*lavaan* package (Rosseel, 2012).

Both Akaike information criterion (AIC) and Bayesian information criterion (BIC) indicated that the bifactor model showed the best fit among the compared four. Root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) were somewhat higher than in previous studies but still indicate a reasonable fit. Bentler's comparative fit index (CFI) and Tucker–Lewis index (TLI) were relatively low. But, as Lin et al. (2021) argue, it is hard to achieve over .90 for a model with many factors with each having many associated items, and indeed, the Chinese version of the Gold-MSI shows similar CFI and TLI values. Overall, our results confirmed that the bifactor structure suggested by the original Gold-MSI outperformed alternative models. Figure 2 presents the factor structure and standardized parameters of the Gold-MSI-J of the bifactor model (Model 2). Please note that the general factor of the original Gold-MSI consists of 18 items while we included all 38 items, because our 18 items with the highest standardized factor loading did not match the English version (such discrepancy is commonly present in other translations of Gold-MSIs). Because all loadings were significant, we decided to keep all of them in the model.

Table 4 presents the summary of the association between demographic information, socioeconomic status, and Gold-MSI-J scores. We used Spearman's ρ, a non-parametric correlation measure, to measure associations between our discrete variables. The age of participants was significantly positively correlated with Perceptual Abilities. Mann–Whitney *U* tests indicated that three subscales of the Gold-MSI-J scores were significantly different between males and females, Perceptual Training, Musical Training, and Singing Abilities. In all cases, females had significantly higher average scores than males (see also Table 5). There were no significant correlations between socioeconomic status and Gold-MSI-J scores. However, because the number of individuals who provided information about their socioeconomic status was rather small, further validation would be desirable for drawing conclusions based on more evidence.

Discussion

This article presents the development and evaluation of the Gold-MSI-J. After a careful translation process, we collected valid responses from 689 Japanese native speakers. The model comparisons confirmed that the internal bifactor structure of the original questionnaire is
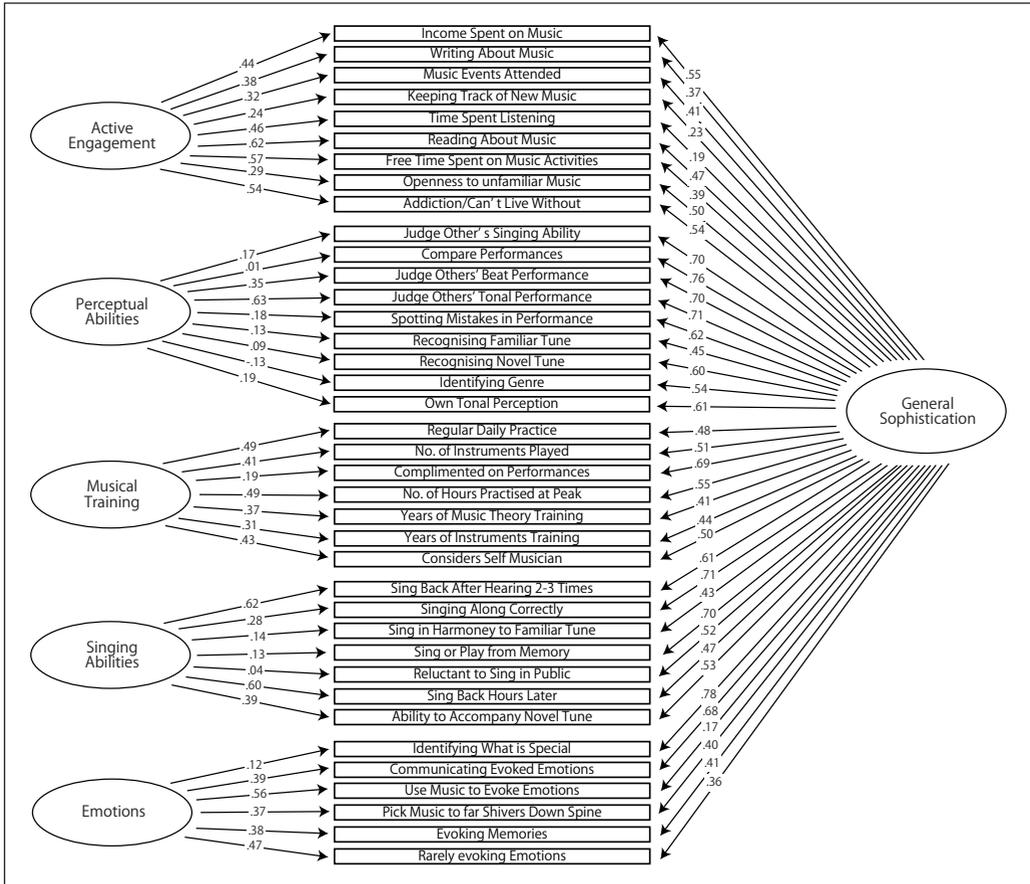
Figure 2. The bifactor model and standardized parameters, applied to the validation data of the Gold-MSI-J.

Table 4. Associations between the Gold-MSI-J and age, gender, socioeconomic status, and education.

	n	Statistics	Active Engagement	Perceptual Abilities	Musical Training	Singing Abilities	Emotions	General Sophistication
Age	689	Spearman's ρ	.03	.19**	.12	.02	–.10	.05
Gender	678	Mann–Whitney *U*	59,240	490,622*	37,930**	48,661*	53,538	49,798
		Rank-Biserial Correlation	.06	–.12	–.32	–.13	.04	.11
SES	142	Spearman's ρ	.07	–.09	.01	–.04	–.09	–.00
Education	496	Spearman's ρ	<.01	–.01	–.01	–.02	–.02	–.02

Note: *n* varies for different items because not all participants provided corresponding answers (no answers or other). The analysis of gender excluded 11 responses that were either "other" (*n* = 5) or "do not want to share" (*n* = 6). Coding used for gender: females = 0 (*n* = 282), males = 1 (*n* = 396). SES: socioeconomic status (income), Education: highest education qualification expected.
*p < .05, **p < .01 (Bonferroni corrected).

Table 5. Breakdown of statistics for the five subscale scores and general sophistication scores for age, gender, socioeconomic status, and education.

Category (n)	Active Engagement			Perceptual Abilities			Musical Training			Singing Abilities			Emotions			General Sophistication		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Gender																		
Female (282)	32.14	8.4	11–56	42.77	9.2	14–62	23.57	9.7	7–46	26.60	8.2	7–46	30.84	5.6	12–42	68.31	18.9	23–114
Male (396)	33.10	9.5	9–59	40.68	9.7	9–63	18.31	8.9	7–49	24.78	8.3	7–47	30.34	5.9	7–47	64.69	19.2	18–120
Age																		
Under 20 (185)	32.37	8.5	11–54	39.56	8.7	16.7–59.4	18.18	8.4	7–43	25.49	7.5	9–43	31.42	5.5	15–41	64.63	17.1	27–103
20–29 (214)	33.48	10.3	9–59	41.08	10.6	9–63	22.01	10.7	7–49	25.83	9.1	7–47	30.71	6.3	10–42	68.01	22	21–120
30–39 (107)	30.50	6.6	15–47	41.87	8.4	19–61	19.41	9.3	7–46	24.63	8.3	7–46	29.68	4.9	7–43	61.97	17.3	18–113
40–49 (96)	31.33	9.1	11–52	42.26	10	14–61.6	20.87	9.5	7–43	24.43	8.6	7–43	28.48	6.2	12–42	64.26	18.9	23–110
50–59 (66)	35.68	8.4	17–53	45.97	7	26–61	22.94	8.5	8–43	27.76	7	12–42	31.64	5.2	15–41	73.09	16	36–111
60–69 (16)	34.25	10.3	11–52	45.81	10.2	21–60	20.50	9.2	8–41	25.56	9.5	10–43	30.25	6.1	16–41	67.69	21.2	23–98
Above 70 (5)	40.20	5.6	36–50	45.40	6	36–52	26.20	12.7	11–42	28.2	3	25–33	32.4	2.8	25–33	79.00	13	63–91
SES																		
No income (29)	28.93	5.8	17–39	40.14	7.4	22–55	27.93	4.6	18–36	16.31	6.4	7–25	22.62	7.2	8–35	57.52	13.3	30–86
<1 m yen (40)	29.33	4.6	19–39	41.13	6.1	26–52	29.33	3.3	21–35	18.43	7.3	7–39	23.4	7.3	12–40	59.78	14.5	36–99
1–3 m yen (51)	30.02	7.4	11–49	39.80	9.6	14–61	29.39	6.4	12–40	18.41	9.1	7–43	22.16	7.9	7–45	58.43	17.8	18–102
3–5 m yen (15)	28.27	6	19–38	38.80	8.2	29–58	25.67	5.4	16–40	16.33	6.8	8–28	22.13	6	12–32	55.07	11.8	34–78
5–7 m yen (13)	32.23	5.7	22–42	38.31	8.4	22–52	26.69	5.5	16–33	18.92	10.1	7–34	21.46	8.4	9–35	59.69	18.8	24–84
Education																		
Second school qualification (26)	31.27	7.3	21–49	40.19	9.5	21.6–53.8	17.19	9	7–37	24.577	8.7	11–42	29.5	5	20–39	63	17.5	33–91
Post 18 vocational course (7)	36.57	8.3	27–48	48.86	6.3	39–60	20.86	8.5	12–35	31	7.4	17–42	31.429	3.3	27–35	76.571	20.5	41–110
Junior college or technical school (8)	33.63	10.4	14–44	44.63	13.9	14–63	24.63	10.5	7–43	29.125	7.5	18–40	32.75	6.3	21–42	74.75	23.3	30–111
Undergraduate degree (317)	33.44	9.1	11–59	41.24	9.3	9–63	20.68	9.5	7–49	25.842	8.1	7–47	31.164	5.6	13–42	67.363	18.6	22–120
Postgraduate degree (138)	32.73	10.4	9–56	41.43	10.3	15–62	20.00	9.9	7–46	25.536	8.4	8–44	30.42	6.4	10–42	66	20.1	21–114

Note: SD: standard deviation; SES: socioeconomic status.
1 million yen = about 7,400 euros, participants from CrowdWorks answered the SES questions.

maintained in the Japanese Gold-MSI. Our results also replicated high internal consistency (McDonald's ω, Cronbach's α, Gutmann's λ6, and inter-factor correlation) and test–retest reliability of the subscales of the questionnaire.

While the relative score structure is reliable and comparable with the original and other translated versions of the Gold-MSI, the comparison of absolute scores seems to show some cultural differences. For example, the original and Traditional Chinese versions demonstrate a higher average score on the five subscales and the general musical sophistication scale than those for the German, Portuguese, and French versions (see Supplemental Table 1). Absolute average scores of the subscales of Gold-MSI-J tend to be on the lower side as compared to other translated versions. It is hard to identify the reasons for such subscale score variability among translated versions because it may stem from nuances in translations but also from some cultural characteristics or differences in the sample composition. With regard to the Japanese results, the modest scores may be partially related to the characteristics of Japanese culture that value modesty and show relatively low self-enhancing motivations (e.g., Heine & Renshaw, 2002; Markus & Kitayama, 1991). Administering objective musical perception and performance tests is an excellent way to identify whether musical abilities are indeed lower in samples from the general Japanese population than in comparable samples from other countries or other factors are contributing to the observed empirical differences. In any case, as Lin et al. (2021) rightly state, direct comparisons of self-report absolute scores across different countries and cultures may not be appropriate due to such potential cross-cultural differences.

The ways we think about musical sophistication are changing continuously. Around the world, there is an increasing awareness that defining musicality in terms of a traditionally used single scale such as years of instrumental training falls short of representing the manifold ways in which people engage with music. It is therefore imperative that researchers around the globe are able to use psychometrically valid tools to measure musical sophistication in its full breadth, not least to encourage cross-cultural studies and to facilitate the comparison of research outputs beyond national and linguistic boundaries. Translating, adapting, and further validating already existing test batteries and self-report inventories, as we have done with the Gold-MSI-J, provide an important step toward reaching these goals.

ORCID iDs

Makiko Sadakata https://orcid.org/0000-0002-3931-0642
Andres von Schnehen https://orcid.org/0000-0002-5232-722X

Supplemental material

Supplemental material for this article is available online.

Note

1. For CrowdWorks participants who received a small monetary compensation (see "Data collection" section for detail), we asked two additional questions using reverse-phrasing for the same content as

Questions 9 and 22. This was done to identify non-serious attempts. (No. 9 original: Pieces of music rarely evoke emotions for me, reversed: Pieces of music often evoke emotions for me. No. 22 original: I spend most of my disposable income on music, reversed: I don't spend much of my disposable income on music.)

References

Aksentijevic, A., Smith, A., & Elliott, M. A. (2014). Rate-specific entrainment of harmonic pitch: Effects of music training. *Music Perception*, *31*(4), 316–322. https://doi.org/10.1525/MP.2014.31.4.316

Chin, T., & Rickard, N. S. (2010). Nonperformance, as well as performance, based music engagement predicts verbal recall. *Music Perception*, *27*(3), 197–208. https://doi.org/10.1525/mp.2010.27.3.197

Collier, G. L., & Collier, J. L. (2007). Studies of tempo using a double timing paradigm. *Music Perception*, *24*(3), 229–245. https://doi.org/10.1525/mp.2007.24.3.229

Degrave, P., & Dedonder, J. (2019). A French translation of the Goldsmiths Musical Sophistication Index, an instrument to assess self-reported musical skills, abilities and behaviours. *Journal of New Music Research*, *48*(2), 138–144. https://doi.org/10.1080/09298215.2018.1499779

Fuji, 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

Furuichi, H., & Umemoto, T. (1975). Bentley ongaku nouryoku tesuto no hyoujun ka [The translation and standardization of the Bentley Measure of Musical Abilities]. *Journal of the Musicological Society of Japan*, *21*, 65–77.

Galton, F. (1883). *Inquiries into human faculty and its development*. Macmillan.

Gelding, R. W., Harrison, P., Silas, S., Johnson, B. W., Thompson, W. F., & Müllensiefen, D. (2021). An efficient and adaptive test of auditory mental imagery. *Psychological Research, 85*(3), 1201–1220. https://doi.org/10.1007/s00426-020-01322-3

Gordon, E. (1967). The musical aptitude profile. *Music Educators Journal, 53*(6), 52–54.

Gembris, H. (1997). Historical phases in the definition of musicality. *Psychomusicology: Music, Mind & Brain*, *16*, 17–25. https://doi.org/10.1037/h0094070

Gordon, E. E. (1995). *Musical Aptitude Profile*. GIA Publications.

Hallam, S. (2010). 21st century conceptions of musical ability. *Psychology of Music*, *38*(3), 308–330. https://doi.org/10.1177/0305735609351922

Hallam, S. (2016). Conceptions of musical understanding. *Research Studies in Music Education*, *38*(2), 133–154. https://doi.org/10.1177/1321103X16671037

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

Harrison, P., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports, 7*(1), 1–18. https://doi.org/10.1038/s41598-017-03586-z

Harrison, P., & Müllensiefen, D. (2018). Development and validation of the computerised adaptive beat alignment test (CA-BAT). *Scientific Reports, 8*(1), 1–19. https://doi.org/10.1038/s41598-018-30318-8

Harrison, P., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods, 51*(2), 663–675. https://doi.org/10.3758/s13428-019-01225-1

Heine, S. J., & Renshaw, K. (2002). Interjudge agreement, self-enhancement, and liking: Cross-cultural divergences. *Personality and Social Psychology Bulletin*, *28*(5), 578–587. https://doi.org/10.1177/0146167202288002

Jimenez, I., & Kuusi, T. (2018). Connecting chord progressions with specific pieces of music. *Psychology of Music*, *46*(5), 716–733. https://doi.org/10.1177/0305735617721638

Kolinsky, R., Cuvelier, H., Goetry, V., Peretz, I., & Morais, J. (2009). Music training facilitates lexical stress processing. *Music Perception*, *26*(3), 235–246. https://doi.org/10.1525/mp.2009.26.3.235

Larrouy-Maestri, P., Harrison, P., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods, 51*(2), 663–675. https://doi.org/10.3758/s13428-019-01225-1

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, H., & Müllensiefen, D. (2020). The Timbre Perception Test (TPT): A new interactive musical assessment tool to measure timbre perception ability. *Attention, Perception, & Psychophysics, 82*(7), 3658–3675. https://doi.org/10.3758/s13414-020-02058-3

Li, J., Wolf, A., Lin, H.-R., & Lothwesen, K. (2021, July). *Validation and replication of the Gold-MSI with a sample of Chinese General* [Paper presentation]. 16th International Conference on Music Perception and Cognition, Virtual hubs in Sheffield, UK; Mexico City, Mexico; Australia; India. https://drive.google.com/file/d/1wkNUOY7BTrc9JOP7oNkxkg9v9_x73buy

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

Lin, H. R., Kopiez, R., Müllensiefen, D., & Wolf, A. (2021). The Chinese version of the Gold-MSI: Adaptation and validation of an inventory for the measurement of musical sophistication in a Taiwanese sample. *Musicae Scientiae, 25*(2), 226–251. https://doi.org/10.1177/1029864919871987

MacGregor, C., & Müllensiefen, D. (2019). The Musical Emotion Discrimination Task: A new measure for assessing the ability to discriminate emotions in music. *Frontiers in Psychology*, https://doi.org/10, 1955. 10.3389/fpsyg.2019.01955.

Madison, G., & Merker, B. (2005). Timing of action during and after synchronization with linearly changing intervals. *Music Perception, 22*(3), 441–459. https://doi.org/10.1525/mp.2005.22.3.441

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Motegi, M., Ogawa, I., & Suzuki, K. (1959). *Tanaka kyouiku kenkyusho ongaku soshitsu shindan tesuto* [Tanaka Educational Institute's Musical Aptitude Judgement Test]. Nihon Bunka Kagakusha Co. Ltd.

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/10.1371/journal.pone.0089642

Nazari, M. A., Ebneabbasi, A., Jalalkamali, H., & Grondin, S. (2018). Time dilation caused by oddball serial position and pitch deviancy: A comparison of musicians and nonmusicians. *Music Perception, 35*(4), 425–436. https://doi.org/10.1525/mp.2018.35.4.425

Ogawa, Y., Murao, T., & Mang, E. H. S. (2008). *Developing a music aptitude test for school children in Asia* [Paper presentation], 10th International Conference for Music Perception and Cognition, Hokkaido, Japan.

Olsen, K. N., Thompson, W. F., & Giblin, I. (2018). Listener expertise enhances intelligibility of vocalizations in death metal music. *Music Perception, 35*(5), 527–539. https://doi.org/10.1525/mp.2018.35.5.527

Ongaku Shinri Kenkyu Sho. (1969). *Onken-shiki Youji no ongaku tokusei tesuto* [Onken's Musical Aptitudes Test for Young Children]. Nihon Bunka Kgakusha Co. Ltd.

Panagiotidi, M., & Samartzi, S. (2013). Time estimation: Musical training and emotional content of stimuli. *Psychology of Music, 41*(5), 620–629. https://doi.org/10.1177/0305735612441737

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

Radvansky, G. A., Fleming, K. J., & Simmons, J. A. (1995). Timbre reliance in nonmusicians' and musicians' memory for melodies. *Music Perception, 13*(2), 127–140. https://doi.org/10.2307/40285691

Rammsayer, T., & Altenmüller, E. (2006). Temporal information processing in musicians and nonmusicians. *Music Perception*, *24*(1), 37–48. https://doi.org/10.1525/mp.2006.24.1.37

Repp, B. H., London, J., & Keller, P. E. (2012). Distortions in reproduction of two-interval rhythms: When the "Attractor Ratio" is not exactly 1:2. *Music Perception*, *30*(2), 205–223. https://doi.org/10.1525/mp.2012.30.2.205

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48, 1–36. http://www.jstatsoft.org/v48/i02/

Rossel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica*, *138*(1), 1–10. https://doi.org/10.1016/j.actpsy.2011.03.007

Satoh, M., Nakase, T., Nagata, K., & Tomimoto, H. (2011). Musical anhedonia: Selective loss of emotional experience in listening to music. *Neurocase*, *17*(5), 410–417. https://doi.org/10.1080/13554794.2010.532139

Schaal, N., Bauer, A.-K., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe. [The Gold-MSI: Replication and validation of a survey instrument for measuring musical sophistication with a German sample.] *Musicae Scientiae*, *18*(4), 423–447. 10.1177/1029864914541851.

Schiavio, A., & Timmers, R. (2016). Motor and audiovisual learning consolidate auditory memory of tonally ambiguous melodies. *Music Perception*, *34*(1), 21–32. https://doi.org/10.1525/mp.2016.34.1.21

Seashore, C. E., Lewis, D., & Saetveit, J. G. (1960). *Seashore Measures of Musical Talents (revised)*. Psychological Corp.

Sloboda, J. A. (1993). Musical ability. In G. R. Bock & K. Ackrill (Eds.), *The origins and development of high ability* (pp. 106–118). Wiley.

Swanwick, K., & Tillman, J. (1986). The sequence of musical development: A study of children's composition. *British Journal of Music Education*, *3*(3), 305–339. https://doi.org/10.1017/S0265051700000814

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

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). 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

Weijkamp, J., & Sadakata, M. (2017). Attention to affective audio-visual information: Comparison between musicians and non-musicians. *Psychology of Music*, *45*(2), 204–215. https://doi.org/10.1177/0305735616654216

Wing, H. D. (1962). A revision of the Wing musical aptitude test. *Journal of Research in Music Education*, *10*(1), 39–46.

Wolpert, R. S. (2000). Attention to key in a nondirected music listening task: Musicians vs. nonmusicians. *Music Perception*, *18*(2), 225–230. https://doi.org/10.2307/40285910

Yu, L., Li, X., Yu, H., Cui, Z., Liao, W., Li, S., Peng, Y., & Wang, Z. (2016). Musicians have larger memory spans for Mandarin tones but not segments. *Psychology of Music*, *44*(5), 1058–1067. https://doi.org/10.1177/0305735615608695

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