MUSICÆ SCIENTIÆ

Article

Musicae Scientiae
1–19

DOI: 10.1177/1029864918811879
journals.sagepub.com/home/msx
SAGE

Examining musical sophistication: A replication and theoretical commentary on the Goldsmiths Musical Sophistication Index

David John Baker
Louisiana State University, USA

Juan Ventura
Louisiana State University, USA

Matthew Calamia
Louisiana State University, USA

Daniel Shanahan
Ohio State University, USA

Emily M. Elliott
Louisiana State University, USA

Abstract

The difficulties associated with measuring the complex construct of musicianship have received considerable attention in the music psychology literature. Multiple measures exist for various constructs, yet the need for the careful replication and documentation of the use of these measures remains an area of critical importance. Here, we describe the replication of the Goldsmiths Musical Sophistication Index (Gold-MSI) in a sample of 346 university students, drawn from both a school of music and a department of psychology. The original approach to modeling the Gold-MSI was followed as closely as possible, and the results replicated well overall. Issues were noted, however, with the characteristics of the sample, the skew of some of the individual items, and the overall use of the bifactor structure. These findings are discussed in relation to the state of measuring musicianship in the current literature, as well as in relation to the larger theoretical concerns surrounding the modeling of complex psychological and musical constructs.

Keywords

Psychometrics, musical sophistication, replication, modeling musicality, confirmatory factor analysis, latent variable modeling

Corresponding author:
David John Baker, School of Music, Louisiana State University, Baton Rouge, LA 70808, USA.
Email: davidjohnbaker1@gmail.com

Measuring musicality is a complex problem. Work from the ethnomusicological (Blacking, 1974) and music education (Murphy, 1999) research community established the challenges that come with attempting to measure musicality and has insisted on the importance of taking culture and context into account with any sort of musical measure. While the nuances of musicality are important considerations at both the individual and comparative level, in order to create viable theories about musicality, the researcher needs to deal with some level of abstraction. One approach is to create abstracted concepts that reduce a complex idea like musicality or musical sophistication into quantitative measures in order to proceed with any sort of modeling (Healy, 2017). Only once some measure of musicality has been defined a priori can researchers begin to make falsifiable claims of how musicality or musical sophistication is related to other human behaviors, such as its relation to cognitive development in children (Jaschke, Honing, & Scherder, 2018), personality (Greenberg, Müllensiefen, Lamb, & Rentfrow, 2015), and even intelligence (Schellenberg, 2004). In this paper we adhere to the term musical sophistication, as the primary analysis presented here focuses on replicating the Goldsmiths Musical Sophistication Index (Gold-MSI) (Müllensiefen, Gingras, Musil, & Stewart, 2014). To foreshadow our discussion, we later claim that music psychologists should use caution in using hypothetical constructs in their modeling of musicality within individuals.[1]

Looking back at the almost 100 years of literature on testing musical sophistication, the field has progressed substantially since Carl Seashore's Measures of Musical Talent was first published in 1919 (Seashore, 1919; Seashore, Lewis, & Saetveit, 1960). Since then, many researchers have developed tools to attempt to measure the individual aspects relating to musicality, brief histories of which can be found in reviews (e.g., Boyle & Radocy, 1987), as well as the litany of tests that have arisen since Seashore's Measures of Musical Talent. Examples of these include the Musical Ear Test (Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010), the Profile of Music Perception Skills (Law & Zentner, 2012), the Ollen Musical Sophistication Index (Ollen, 2006), the Bentley Test of Musical Ability (Bentley, 1966), Gordon's Advanced Measures of Musical Audiation (Gordon, 1989), the Wing Standardized Tests of Musical Intelligence (Wing, 1968), the Gold-MSI (Müllensiefen et al., 2014), the MUSE (Chin & Rickard, 2012) and MUSEBAQ (Chin, Coutinho, Scherer, & Rickard, 2018), as well as the Musical Ear Training Assessment (Wolf & Kopiez, 2018). Each of these test batteries produces a continuous variable, implying that each trait related to musical perception is measured and conceptualized as being on a continuum, rather than a categorical, dichotomous variable often reflected in the colloquial language used to talk about who is or is not a musician. Recent calls have even been made in music psychology journals that provide a clear and compelling rationale for the research community to stop dichotomizing musicianship (Daly & Hall, 2018), especially considering the amount of tools that exist to measure constructs of interest continuously.

We note that some tools like the Gold-MSI and the MUSEBAQ use latent variable modeling (Loehlin & Beaujean, 2016) to accomplish their goals, suggesting that the concept of musicality might be akin to other reliably measured constructs such as intelligence. At the conceptual level, some of the tests even label themselves a model of musical intelligence (e.g., Bentley, 1966), similar to how Seashore may have thought of musicianship as a latent ability, innate within the individual (Seashore, 1915, p. 129).

Despite this variety and granularity of measurement, researchers sometimes revert back to the arbitrary dichotomy that many of these tools were developed to avoid (e.g. musicians/non-musicians). The most recent—and possibly clearest—example of this is a meta-analysis looking at musicians and memory, published by Talamini and colleagues (Talamini, Altoè, Carretti, & Grassi, 2017). In conducting their meta-analysis, they used

university music education at the undergraduate level as the criterion to decide who did and did not qualify as a musician. The authors summarized studies and largely concluded that musicians often outperformed their non-musician counterparts on multiple measures of memory. The findings suggested that musical training might have something to do with this; they give a few reasons for their findings. While the relationship between musicianship and memory is interesting in its own right, the authors note when reflecting on the limitations of their study that,

> There is currently no standard for describing the characteristics of musicians and nonmusicians, and several potentially interesting characteristics are very often not reported (e.g., hours of daily practice, instrument played, etc.). There are examples in the literature of questionnaires that can be used to draw up a complete profile of participants (both musicians and nonmusicians, [86, 87]) and, since most studies comparing musicians with nonmusicians are quasi experimental, a thorough description of the two groups would be of fundamental importance. In many circumstances, a shortage of information makes it impossible to disentangle whether or not musicians' enhanced performance is an effect of their music training. Studies also often failed to report or control for variables that might explain the difference between groups: for example, not all the studies analyzed here controlled for general cognitive abilities (e.g., intelligence) (Talamini et al., 2017).

In their citations, they describe both the MUSE (Chin & Rickard, 2012) and the Gold-MSI (Müllensiefen et al., 2014) as questionnaires developed to pick up on their aforementioned characteristics such as hours of daily practice and instrument played. While they do mention the Gold-MSI as an example of a questionnaire that addresses this problem of having inconsistent methods of measuring the characteristics of musicians and non-musicians, other facets of the tool such as its incorporation of objective tests and its reliance on a bifactor structure (5 + 1) that explains variance at higher and lower levels are worth noting here. This design reflects the increasingly adopted view of musicianship as being polymorphic (Levitin, 2012) and reflects the diversity of musicianship often advocated for, as discussed above (Blacking, 1974; Murphy, 1999).

While there are plenty of tools that have been developed in order to measure various permutations of musical sophistication, we now turn our attention fully to the Gold-MSI. The Gold-MSI has been rapidly adopted by the music psychology community: at the time of writing, the original Gold-MSI paper has been cited in 226 different studies, according to Google Scholar. Developed in a multi-step process, as detailed in Müllensiefen et al. (2014), the test is composed of a self-report questionnaire inventory with five sub-scales, two objective listening tests (a melodic memory task and beat perception task), as well as a sound similarity sorting exercise. The test has also been translated into other languages including German (Schaal, Bauer, & Müllensiefen, 2014) and Chinese (Lin, Kopiez, Müllensiefen, & Wolf, 2018). Recently the test has been modified and adapted using more sophisticated statistical modeling, resulting in shorter versions of the test, which reportedly have the same construct validity as the original (Harrison, Collins, & Müllensiefen, 2017). In its complete form, the Gold-MSI takes about 30 minutes to complete, although some researchers elect to only use relevant subcomponents such as solely the Musical Training sub-scale (Baker & Müllensiefen, 2017; Farrugia, Jakubowski, Cusack, & Stewart, 2015). The tool has proved especially useful in studies where it has been combined with other measures such as measures of personality (Greenberg et al., 2015) and socio-economic status (Müllensiefen et al., 2014), and has been used as a tool to measure aspects of development (Müllensiefen, Harrison, Caprini, & Fancourt, 2015).

The Gold-MSI has succeeded in what it set out to do: to establish a validated psychometric tool that captures a polymorphic view of musicianship. The tool is set up to catch that sense of musicianship from individuals with musical backgrounds as diverse as the Internet disc jockey, the ballerina, and the classical music newspaper critic, none of whom may have had an instrumental musical lesson in their lives, yet are steeped in music daily and would be discounted from many of the studies undertaken in music psychology. This diversity in musical ability is helpful to capture, but the nebulous and debated nature of musical sophistication on a theoretical level can lead to discrepancies at the modeling level.

Given its increased use, investigating the validity and reliability of the Gold-MSI in various settings deserves a particularly high degree of attention. In this paper, we follow the recent wave of replication research happening in the field of psychology (Open Science Collaboration, 2015) in order to provide further evidence for the replicability of our psychometric tools (Frieler et al., 2013; Ioannidis, 2005).

Our goals in this paper are to examine the fit indexes of the Gold-MSI in a sample of North American university students that is reflective of many of the sample demographics found in psychology journals today (Henrich, Heine, & Norenzayan, 2010), to provide a detailed analysis and commentary of the item level and sub-scale item level, and reflect on the use of bifactor latent variable models in music psychology. We believe that replicating the original findings is necessary given that most administrators of the Gold-MSI will not be operating with the scale of sample used in the original paper ($N = 147{,}636$); additionally, with a large amount of studies being conducted within a university environment, researchers should be aware of any sort of bias that could emerge when they do not have access to as large and diverse a sample as the original paper.

Method

Data for this project were collected as part of a larger study exploring the relationship between working memory, general intelligence, and musical sophistication that ran from February 2016 through November 2017. This study was approved by the Louisiana State University Institutional Review Board.

Participants

Participants for this study were drawn from the student population at Louisiana State University (LSU). Subject pools from both the School of Music and psychology department were used: students undertaking either an undergraduate-level music theory or psychology course were given the option to fulfill part of a required research component by participating in the study. Participants were also recruited via an online platform to access a broader range of subjects. Additionally, we recruited more subjects with higher degrees of musical training to ensure a diverse range of musicality. A subset of musicians was paid USD 15 for their time. The total sample ($N = 346$, $M_{Age} = 20.17$, $SD = 2.72$) consisted of 217 women (63%), 129 men (37%), and 5 (1%) participants who did not disclose their gender.

Procedure

Participants completed a battery of tests beginning with the self-report measures used in the Gold-MSI (Müllensiefen et al., 2014), a novel tone span task currently in development

(Ventura, 2018), followed by operation and symmetry span tasks (Unsworth, Heitz, Schrock, & Engle, 2005), three objective listening tasks incorporated in the Gold-MSI (beat perception, melodic memory, sound similarity), and finally Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998) as a measure of general fluid intelligence. All participants were run individually in the lab while seated in front of a desktop computer. The entire experiment lasted approximately 90 minutes.

Materials and stimuli

The Gold-MSI was administered with a browser-based implementation of the survey designed by the first author using the jsPsych package (De Leeuw, 2015). The operation, symmetry, and tone span tasks, along with the Raven's task were all run in E-Prime 2.0 with Windows 7 (Schneider, Eschman, & Zuccolotto, 2002). The objective listening tests from the Gold-MSI were administered using version 1.0 of the tests in Qualtrics, through the help of Daniel Müllensiefen. The melodic memory test required participants to identify whether two consecutively played transposed melodies had the same pitch structure, with a binary yes/no response along with a three-point Likert scale confidence rating of the judgment. The beat perception test required participants to listen to short excerpts that included a beep click-track that was either congruent or incrementally offset to the recording, and to indicate whether the clicks were in-time or not, along with giving a three-point confidence rating.

The other measures (complex span tasks and measures of general fluid intelligence) were not the focus of the current work, and were therefore excluded from these data analyses.[2] Participants listened to all musical stimuli on high-quality Sennheiser headphones in a quiet room with the volume adjusted to a comfortable level, as determined by the researcher. Post experiment, participants were made aware of the objectives of the research and given an opportunity to sign up for further studies.

Results

Data from this experiment were analyzed as closely in accordance to the original Gold-MSI as possible (Müllensiefen et al., 2014).[3] Following the original Müllensiefen et al. paper, data were analyzed using the statistical computing software *R* (R Core Team, 2017) using the *lavaan* (Rosseel, 2012) and the *psych* (Revelle, 2014) package.

Data cleaning

Before running any analyses, data were inspected for both univariate and multivariate outliers. Although no univariate outliers were detected due to the use of Likert scales, 18 outliers were identified by identifying participants that exceeded a $p < .001$ significance Mahalanobis distance (a standard metric for multivariate outliers) with degrees of freedom equal to the number of variables (Tabachnick, Fidell, & Osterlind, 2001). Regression models predicting both objective listening test measures were run using data sets with and without the multivariate outliers. Differences in R^2-values were deemed to be negligible (.009) thus the 18 data points were included in our analyses.

Inspecting our data, we additionally noticed a negative skew to the distribution of the General factor, as can be seen in in Figure 1. This skew is most likely due to the subject pools used in the recruitment process. Regardless, we proceeded with the analyses due to the similar

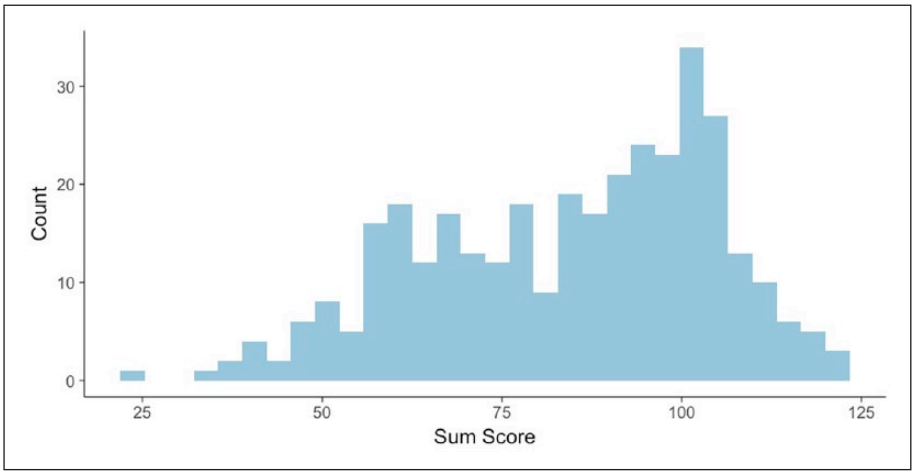

Figure 1. Distribution of scores on the General factor in Louisiana State University sample.

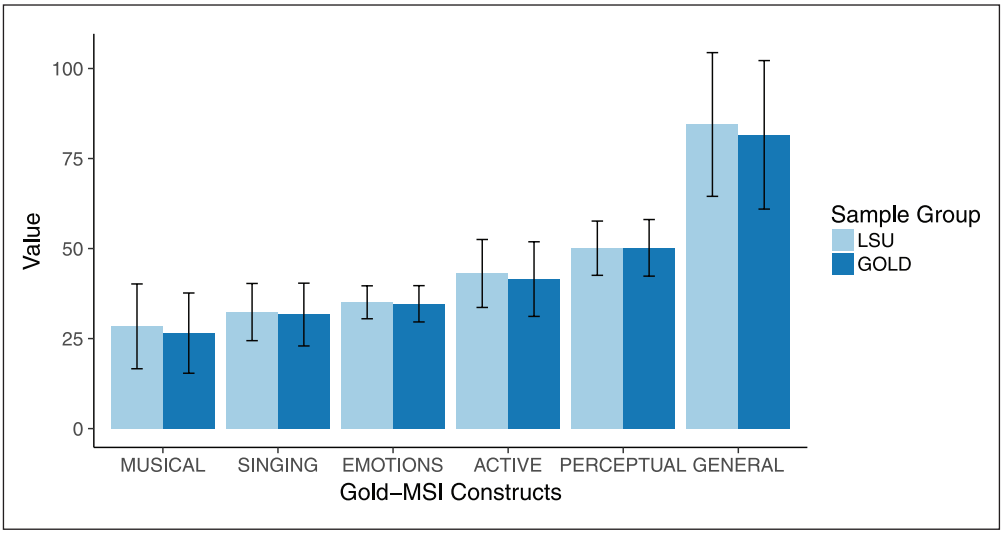

Figure 2. Mean and standard deviation of Gold-MSI constructs in both samples.

means and standard deviations apparent in Figure 2, all of which would be expected in a direct replication.

Descriptive Statistics

Mean sub-scale data from the present study as compared to the original Gold-MSI paper are presented in Figure 2. Item-level data for the LSU sample can be found in Table 1 and Figure 3.

All individual items were then screened for skewness and kurtosis using the *psych* package in *R* (Revelle, 2014). Though we believe that our sample size was large enough to overcome difficulties with skew and kurtosis (Tabachnick et al., 2001), visually inspecting items revealed

Table 1. Item-level descriptive statistics for LSU sample including Cronbach's alpha.

Sub-scale	LSU/Gold α	Question	*M*	SD	Skew	Kurtosis
Active Engagement	.84/.87	Freetime	5.34	1.63	−0.88	−0.07
		Writing	3.30	1.68	0.40	−0.77
		MusicalStyles	5.45	1.34	−1.11	1.45
		SearchInternet	5.30	1.52	−0.88	0.14
		SpendMoney	3.95	1.88	−0.04	−1.25
		Addiction	5.74	1.43	−1.26	1.14
		KeepTrack	5.53	1.21	−1.08	1.30
		LiveEvents	4.33	1.77	−0.09	−0.94
		ListenAttentively	4.14	1.72	0.17	−0.90
Perceptual	.83/.87	Singer	5.80	1.19	−1.27	1.97
		HearFirstTime	6.20	0.81	−1.09	1.71
		HardSpot	5.09	1.53	−0.72	−0.32
		ComparePerf	5.37	1.46	−1.17	0.92
		SameSong	5.53	1.09	−0.94	1.05
		HearBeat	6.04	1.18	−1.65	2.96
		HearTune	5.78	1.17	−1.09	1.05
		SelfTonal	4.95	1.79	−0.65	−0.71
		IDgenre	5.35	1.12	−0.98	0.88
Musical Training	.90/.90	NeverComplimented	5.01	2.14	−0.79	−0.91
		NotConsiderSelf	4.46	2.46	−0.31	−1.61
		RegularPractice	4.53	2.23	−0.49	−1.30
		PeakInterest	4.17	1.94	−0.45	−1.07
		MusicTheory	2.94	2.16	0.64	−1.11
		Formal	4.41	2.12	−0.56	−1.09
		NoInstruments	2.86	1.73	0.87	−0.17
Singing Ability	.79/.87	JoinIn	4.49	1.70	−0.44	−0.70
		SingByMemory	5.66	1.55	−1.32	1.10
		HitRightNoteSingAlong	4.71	1.98	−0.59	−0.90
		SinginHarmony	4.54	1.81	−0.36	−1.02
		DontSingPublic	4.15	2.00	−0.11	−1.30
		SingBack23	5.26	1.31	−1.03	0.93
		HearOnceSingBack	3.55	1.54	0.13	−0.91
Emotion	.72/.79	ChooseMusic	5.38	1.57	−1.14	0.70
		PiecesEmotion	6.04	1.19	−1.71	3.32
		ExciteMotivate	6.34	0.87	−2.17	7.66
		IdentifySpecial	5.01	1.34	−0.68	0.07
		TalkEmotionsPiece	5.82	1.15	−1.48	3.07
		EvokesMemories	6.48	0.81	−2.22	8.27
General Sophistication	.90/.93					

Note. LSU = Louisiana State University.

inconsistent distributions across many of the items. Figure 3 exhibits the distributions of each of the items reported in Table 1. Although these data were not reported in the original paper (Müllensiefen et al., 2014), we noted a number of items that did not exhibit normal distributions. Table 2 presents the initial correlations between the self-report scales and the objective

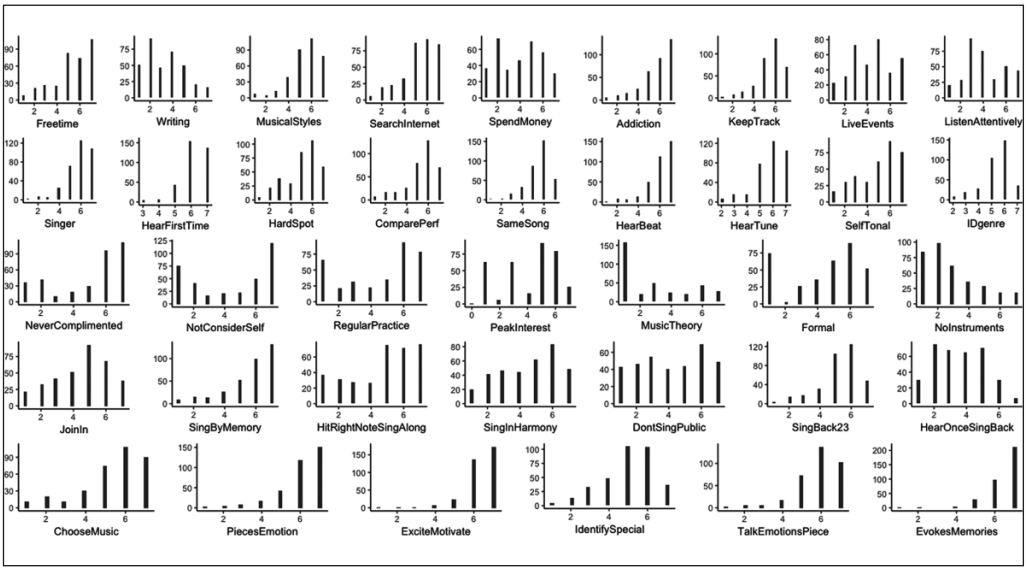

Figure 3. Item-level data.

tests in our sample. We note here that the general-level demographics from this study are dissimilar to those reported in the original paper. The original paper's mean age was much higher ($M_{Age} = 32.2$, $SD = 15$) compared with ($M_{Age} = 20.17$, $SD = 2.72$), the original paper contained more men than women (54.7%) compared with the current sample (37%), and had a much larger spread of education ranging beyond solely a university student population.

Descriptive statistics for listening tests

In our sample, the mean accuracy for the melodic memory task was $M = .64$ ($SD = .15$) and for beat perception mean accuracy was $M = .68$ ($SD = .13$). Our mean accuracy ratings were lower in comparison to the accuracy rates from the original paper for melodic memory $M = .75$ ($SD = .17$) and beat perception $M = .77$ ($SD = .16$).

Hierarchical factor structure

Following Müllensiefen et al. (2014), we ran a Schmid–Leiman solution (Schmid & Leiman, 1957) on our data set in order to replicate the initial findings using the maximum likelihood estimation with robust (Huber–White) standard errors and a scaled test statistic that is (asymptotically) equal to the Yuan–Bentler test statistic in order to compensate for the high skewness of the items, as described above. As noted in Müllensiefen et al. (2014), the Schmid–Leiman is a variation on a hierarchical model where loadings from the General factor are separated from the group factor. In this type of model the General factor and the 38 individual items covary directly. In total, this model has 114 free parameters to estimate. Our model values are presented alongside the initially reported values. We note that in our model, all 38 items load on the General factor in addition to loading on exactly one of the group factors. We state this explicitly due to the fact that the original paper only depicts the 18 items loading on the General factor (Müllensiefen et al., 2014). Comparison of model parameters can be seen in Table 3 and estimates of the factor

Table 2. Correlations between Gold-MSI sub-scales and objective tests.

LSU Correlations	General	Active Engagement	Perceptual Abilities	Emotions	Singing Ability	Musical Training
General		.73	.79	.65	.78	.88
Active Engagement			.55	.71	.50	.56
Perceptual Abilities				.60	.65	.64
Emotions					.45	.49
Singing Ability						.59
Musical Training						
Melodic Memory	.24	.14	.18	.18	.19	.26
Beat Perception	.33	.20	.29	.19	.25	.36

Table 3. Schmid–Leiman model fits of LSU sample and original paper.

Statistic	Gold	LSU
N	147,636	346
χ^2	166,170	1290
df	627	627
BIC	167,448	43,621
TLI	.87	.87
CFI	.88	.87
RMSEA	.06	.06
SRMR	.06	.06

Note. df: degrees of freedom; BIC: Bayesian Information Criterion; TLI: Tucker-Lewis Index; CFI: Comparative Fit Index; RMSEA: Root Mean Square Error of Approximation; SRMR: Standardized Root Mean Square Residual.

loadings of the model can be seen in Table 4. Modeling fitting was undertaken using the *lavaan* cfa() function by assigning each item to its respective sub-scale, and then assigning all items to a General factor. The analysis was run on the original item-level data, allowing factors to be orthogonal, and used the maximum likelihood estimation with the same details as described above.

Structural equation models with objective tests

We next fit structural equation models based on the correlations between the five sub-scales of the Gold-MSI and the two objective listening measures. Following the Müllensiefen et al. (2014) paper, we did not include the non-significant parameter estimates from the Active Engagement and Emotion sub-scales to beat perception. A comparison of the two studies' fit indexes can be found in Table 5. Finally, we fit a structural equation model based on the General factor summed scores with the two objective listening tests shown in Table 6. The structural equation model was run with the sem() function in *lavaan* by first defining each sub-scale with the related items, then explicitly defining the regressors and covariates in the model.

Relation to objective tests

In our structural equation modeling we similarly found that the correlation using the melodic accuracy and beat perception scores between the two listening sub-tests in the present study ($r = .08$) to

Table 4. Structural equation model replication of Figure 1 from original Gold-MSI paper.

Sub-scale	Item	LSU	Gold			LSU	Gold
Active Engagement	Income Spent on Music	.63	.75	General Sophistication	Writing about Music	.86	.88
	Writing about Music	.72	.69		Reading about Music	.86	.92
	Music Events Attended	.51	.52		Free Time Spent on Music Activities	1.13	.97
	Keeping Track of New Music	.80	.72		Addiction/Can't Live Without	.68	.94
	Time Spent Listening	.88	.97		Compare Performances	1.04	.92
	Reading about Music	.76	.76		Own Tonal Perception	1.20	.96
	Free Time Spent on Music Activities	.53	.60		Regular Daily Practice	1.60	1.30
	Openness to Unfamiliar Music	.45	.34		No. of Instruments Played	1.03	.97
	Addiction/Can't Live Without	.74	.77		Complimented on Performances	1.59	1.30
Perceptual Ability	Judge Others' Singing Ability	.32	.41		No. of Hours Practised at Peak	1.60	1.11
	Compare Performances	.30	.04		Considers Self Musician	1.95	1.40
	Judge Others' Beat Performance	.50	.49		Sing Back after Hearing 2–3 Times	.35	.91
	Judge Others' Tonal Performance	.16	.63		Singing Along Correctly	.49	1.08
	Spotting Mistakes in Performance	.37	.40		Sing in Harmony to Familiar Tune	.96	.91
	Recognising Familiar Tune	.46	.19		Sing or Play from Memory	1.01	1.10
	Recognising Novel Tune	.52	.21		Reluctant to Sing in Public	.76	.94
	Identify Genre	.44	.10		Ability to Accompany Novel Tune	.49	.90
	Own Tonal Perception	.18	.42		Identifying What is Special	.86	.93
Musical Training	Regular Daily Practice	.44	1.57				
	No. of Instruments Played	.50	.82				
	Complimented on Performances	1.27	.72				
	No. of Hours Practised at Peak	.83	.71				
	Years of Music Theory Training	.96	1.43				
	Years of Instrument Training	.32	1.67				
	Considers Self Musician	.51	.90				

(Continued)

Table 4. (Continued)

Sub-scale	Item	LSU	Gold	LSU	Gold
Singing Abilities	Sing Back after Hearing 2–3 Times	.65	.53		
	Singing Along Correctly	.28	.87		
	Sing in Harmony to Familiar Tune	.16	.81		
	Sing or Play from Memory	.29	.48		
	Reluctant to Sing in Public	.35	.88		
	Sing Back Hours Later	.98	.52		
	Ability to Accompany Novel Tune	1.03	.58		
Emotions	Identifying What is Special	.45	.18		
	Communicating Evoked Emotions	.50	.54		
	Use Music to Evoke Emotions	.31	.50		
	Pick Music for Shivers Down Spine	.38	.53		
	Evoking Memories	.71	.48		
	Rarely Evoking Emotions	.30	.53		

Table 5. Structural equation model replication of Figure 2 from original Gold-MSI paper.

Factor Loadings				
	Manifest Variables	Latent Variables	LSU	Gold
	Melodic Memory	Active	−.06	−.11
		Perceptual	−.05	.15
		Emotions	.07	−.01
		Singing	.03	.07
		Musical	.05	.23
	Beat Perception	Perceptual	.02	.19
		Singing	−.01	.02
		Musical	.04	.23
Correlations				
	Active	Perceptual	.66	.56
		Emotions	.91	.66
		Singing	.66	.49
		Musical	.65	.46
	Perceptual	Emotions	.75	.63
		Singing	.87	.71
		Musical	.74	.58
	Emotions	Singing	.65	.48
		Musical	.59	.38
	Singing	Musical	.76	.63

Table 6. Structural equation model replication of Figure 3 from original Gold-MSI paper.

Exogenous Variable	Endogenous Variables	LSU	Gold
General Musical Sophistication	Beat Perception	.35	.37
	Melodic Memory	.25	.28
Manifest Variables Correlations		.08	.16

Note. LSU = Louisiana State University.

be of a lower magnitude than the original paper ($r = .16$). Our results corroborate the initial claims of Müllensiefen et al. (2014) that these tests are not measuring the same construct.

General discussion

Model fits

In this paper, we have examined the extent to which the Gold-MSI functions as a stable psychometric tool. In order to do this, we used a sample of $N = 346$ students from a large North American university in southern United States, representative of the samples of a large amount of psychological research (Henrich et al., 2010).

The overall scores yielded from the Gold-MSI closely match the means and standard deviations reported in the original paper (see Figure 2). These similar descriptive statistics between the two samples occurred despite problems at the item level with our data, displayed in Figure 3. For example, many of the items from the tests do not exhibit a normal distribution centering on the

middle of the seven-point scale, and in our data set some items show an alarming degree of skew and kurtosis. Specifically, many of the items on the Emotions sub-scale demonstrated a negative skew and did not behave like the other items on the Index. The negative skew of our sample is also reflected in the Musical Training sub-scale in items 19–25. Considering the two subject pools the study draws from, these results reflect that we sampled from at least one of the extremes of musical sophistication that the Index was intended to measure. We believe it important to note that as the Gold-MSI was originally developed to measure musical sophistication in the general population, researchers need to ensure that the population they are sampling from indeed reflects that of the general population and not just a university's subject pool. Additionally, we believe that the skewing of items in the Emotion sub-scale might also be reflective of the younger demographics present in our study; relatively younger people often report music as being more important to their lives and identity than older people (Tarrant, North, & Hargreaves, 2002).

Examining the degree to which the item-level data were able to recreate the factor structure underlying the Gold-MSI, we turn our attention to tables 3 and 4. After running the Schmid–Leiman analysis on our data, we were largely able to replicate the model fits, as evident in Table 3. However, the fit statistics did not reach the preferred level of a comparative fit index (CFI) above .9, as recommended in the literature (Beaujean, 2014; Hu & Bentler, 1999; Kline, 2015). This is in contrast to other literature that reports this level of fit is perhaps to be expected with complex, hierarchical or bifactor models (Marsh, Hau, & Grayson, 2005; Marsh et al., 2010). Inspecting Table 4 more closely, we report generally close factor loadings on the items, which suggest stable replications. Analyzing differences between item-level discrepancies at this level would be entirely post hoc and susceptible to type I errors. We do, however, highlight the differences on factor loadings in Table 5 between the Musical Training latent variable and the degree to which it is predictive in both samples. In the original paper, the loading relatively dominated the others by a factor of over 10 in some cases, whereas in our sample there is less variability on the loadings. Lastly, turning to Table 6, we found similar factor loadings between the General Sophistication latent variable and both of the objective tests, which again is indicative of a successful replication. The correlation between melodic memory and beat perception was lower in our sample ($r = .08$) than the original paper ($r = .16$).

Considering both the descriptive statistics, as well as the close proximity in fit indexes as shown in Table 2, we can say with a fair degree of confidence that the Gold-MSI predictably replicated under different experimental conditions. Both the means and standard deviations of the items map similarly onto the original sample, giving credence to idea that the measurements are meaningful from study to study.

As for the factor structures, although the factor loadings did not match entirely as we might have expected, the fit indexes for each originally reported measure generally tended to be in the relative magnitude and direction of the original loadings. One point worth noting here is the clear display and strength of the factor loadings presented in Table 3 that load from the General factor onto items that traditionally deal with music performance. Referring to Table 3, of the 18 items that originally loaded onto the General Sophistication variable, 10 of them draw from both the Musical Training and Singing Ability sub-scales. While the highest factor loadings in our sample come from the Musical Training sub-scale, it brings to light a methodological question: If the General factor is meant to capture the diversity of ways in which someone can be musical, then why are the majority of the items from sub-scales traditionally associated with the exact types of activities the Gold-MSI was trying to step beyond? The authors note in the original paper that,

> for the practical purposes of the development of a new inventory of musical sophistication, the difference in fit between Models 1, 2, and 4 has no consequences, except for the construction of a general scale of musical sophistication indexing the general factor (Müllensiefen et al., 2014, p. 6).

The authors then go on to describe that the way they determined which items should load on the original factor was to order the factor loadings of the model and then, examining them visually with a scree plot, they deemed any items with values above the 0.88 cluster as belonging to the General factor. Although this factor does draw from all scales at least once, the authors note that there is "a clear preponderance of items from the Musical Training and the Singing Abilities sub-scales" (Müllensiefen et al., 2014, p. 6).

While this might not seem like that pressing of an issue, it may cause some alarm that the Gold-MSI is still picking up a large degree of association from musical activity that is often associated with performative musical culture. This clear preponderance of items related to performance is not necessarily a problem, but we would assert that were a General Sophistication structure in a bifactor model to reflect a more domain-general latent ability, then one would expect the items to be more evenly distributed across the five other sub-scales. This of course is a complex, theoretical question of the degree to which a latent variable model succeeds in partitioning unique subscale variance within covariance frameworks. This complexity is also reflected in that the Gold-MSI is a complex confirmatory factor analysis model with more than five factors, with some factors having over five items on each factor. According to other methodological papers (Marsh et al., 2005; Marsh et al., 2010), models with this degree of complexity have a low chance of exceeding relative fit indices (CFI, Tucker–Lewis Index) of greater than .9 and that researchers should be more cautious in interpreting cut-off criteria for the fit indices (Irwing & Hughes, 2018). Given the complexity of these models, further research might also consider exploring alternative modeling choices such as exploratory bifactor structural equation modeling (Morin, Arens, & Marsh, 2016). Although using such a complex model captures much of the elusive, continuous scale of musicianship that is needed as a field—especially given the lack of standardization of measurement in music psychology—we suggest that recent insights from the field of cognitive psychology (Kovacs & Conway, 2016) might provide a more interpretable framework going forward for capturing the types of musical activity that is of interest to music psychologists.

Beyond model fits

At this point the discussion might revert to square one by defining musicianship according to first principles, although insight from recent work in cognitive psychology might offer an alternate route forward in the discussion. Drawing from recent developments in cognitive psychology, there has been debate over the measurement of the idea of *g*, a hypothetical construct thought to be representative of intelligence (Ritchie, 2015) that is also derived from various forms of factor analysis. While extremely useful as a construct, Kovacs and Conway have recently argued that psychologists should begin to move away from the idea of *g* as being causally responsible for the positive manifold[4] (Kovacs & Conway, 2016). Framing their argument by talking about the practical implications of an ontologically real intelligence that could be located *in* the brain, the authors note,

> Therefore, if the concept of general intelligence is correct, then the following statement is valid: "John used his general intelligence to correctly answer items on both the vocabulary test and the mental rotation task." This however is substantially different from the statement: "If John performs better on the vocabulary test than most people, it is likely he will perform better on the mental rotation test as well," because the latter statement leaves the possibility open that John in fact did not use the same general cognitive ability to solve items in the vocabulary test and the mental rotation test, respectively (p. 154).

If one were to transfer this logic here and go a step further, knowing that the statistical tools used to create constructs like *g* are the same as those used to model the General factor of

the Gold-MSI (various factor analyses), if the concept of musical sophistication is correct, then the following statement would also be valid: "Mary used her musical sophistication to correctly answer items on both the melodic memory and beat perception task." Following the logic, we can further assert that there is a meaningful difference in saying "If Mary performs better on the melodic memory test than most people, it is likely she will perform better on the beat perception test," which again shows that there is a possibility that there are different cognitive processes at work for both of these tasks. We see evidence of this in the data in Table 6. There is a low correlation between the two objective tests when modeling the degree to which the General factor of the Gold-MSI predicts performance on the objective tasks. This point was also highlighted in the original paper.

What we hope to show in this parallel is that to understand either *g* or any of the composite scales of the Gold-MSI as any sort of ontological, causal reality that explains human behavior in a satisfactory way, we would be confusing a statistical abstraction for an underlying process. This is not a novel claim, and we explicitly note that arguments for a modular (Peretz, 2006), polymorphic (Levitin, 2012) view of musicianship inspired the creation of tools like the Gold-MSI. We remind readers that the composite numbers created by the psychometric tools used in music psychology only help us talk about the highly complex underlying processes. With respect to music, the separated underlying processes are related in that they enable our musical activities. They are not unified, however, as a monolithic internal resource. Asking individuals many questions about their musical activity, given the nature of the questions, could give rise to music psychology's own positive manifold. This latent variable would never be some sort of internal resource that an individual would draw from, but, in every situation, would be a collection of separate but related abilities used to complete a very specific task related to a very small subset of activities in the entire universe that would comprise all musical abilities.

Our claim here is not to abandon these tools, and we argue that they are useful; without these tools it would be even more difficult to find unifying threads in the music psychology literature. To have established the relationship between a score on a psychometric test and other neurological and behavioral constructs is significant. We only point out that the causal element driving this relationship, if thought of as being "musicianship" as operationalized above, would be poorly understood.

We instead argue for two points for the music psychology community to consider. The first is that cognitive factors such as working memory capacity, general fluid intelligence, and other confounding variables like socio-economic status should be standard in conducting studies on music perception. Although the variance accounted for in some of these studies is relatively small, to have a domain-general construct such as working memory capacity contribute significantly in wide-ranging studies from jazz (Nichols, Wöllner, & Halpern, 2018), to sight-reading (Meinz & Hambrick, 2010), to responses to tapping along to expressive timing (Colley, Keller, & Halpern, 2017), suggests that something beyond musical sophistication is at play. Similar evidence for significant correlations between musical ability, general intelligence, and socio-economic status can be found in Swaminathan, Schellenberg, and Khalil (2017) and Corrigall, Schellenberg, and Misura (2013), respectively. Other factors related to baseline cognitive ability such as crystalized intelligence might also be relevant given certain hypotheses, but we did not include such factors as measures such as working memory capacity and general fluid intelligence by definition should not be fully dependent on any sort of long-term memory mechanisms.

The second point is that researchers should investigate the degree to which specific predictors that are not bundled as latent variables are able to predict outcome variables. While this approach is arguably more effortful, rather than grouping many indicator variables as forming a latent variable, less noise would be introduced into models of cause and effect. Another future line of research might be to consider re-examining past literature on questions that have been used in research on music psychology, seeing which items have successfully predicted variables of interest, then

developing a comprehensive self-report survey that is treated as a catalog of separate items, rather than one composite latent trait.

Considering the second point raised, specific variables that relate to a musical task might be able to better predict performance on a musical task. Treating musicality as a related but not unified processes would hopefully then lead to better modeling of the causal mechanisms that lead to variation at the individual level. To illustrate this point, we ask the reader to consider entertaining a thought experiment where musical sophistication could experience its own Flynn Effect.[5] Over the course of time, would individuals become more musical as a collective? Imagine retrospectively administering any of the aforementioned musical psychometric tests a century ago when both Seashore and Charles Spearman (Spearman, 1904) would have been first testing out their tools. Take, for example, surveying someone with items from the Gold-MSI relating to musical engagement, especially items dealing with tracking music on the Internet or listening actively to music via some device. By the nature of technology, individuals from a century ago would score lower on those items leading to a lower composite General factor score on their musical sophistication. Or, if the ability to encode and remember melodies is largely dependent on the degree of exposure to patterns in a culture's melody via the mechanism of statistical learning (Saffran, Johnson, Aslin, & Newport, 1999), would the types of melodies, as well as the degree of music an individual was exposed to on a daily basis diminish an individual's ability to score highly on a melodic memory task? We argue that an affirmative answer to either of the two scenarios provides a rationale for researchers to consider selecting more specific item-level predictors for models of music perception that reflect casual mechanisms that have more ontological variety. Lastly, not related to experimental integrity, but more anecdotally, if we were to move further from a unified notion of musicality in educational contexts, one could imagine outcomes where an individual was either self- (or other) diagnosed as "not being musical enough". In this case, the individual in mind would not be understood as having a lack of an internal resource, but rather a lack of some other ability, be it cognitive or experiential.

Conclusions

We have replicated the results of the Gold-MSI to a satisfactory degree. Comparing the Gold-MSI's relative fit measures to some of the modeling literature, we note that the model fits are not conclusively good and suggest that the reason might stem from both model complexity and possibly a deeper problem of what music psychometric tools purport to measure. While caution is needed in interpreting such complex models, we have provided a detailed summary of quirks that researchers might run into when using the Gold-MSI and, additionally, put forward arguments for music psychology to consider when using latent variables in their statistical modeling. We suggest that the music psychology community should not abandon the use of standardized psychometric tools for more unified research, but as a research community should consider ways more specific theories might better explain trends in music perception. Marking the upcoming centenary of Carl Seashore's Measures of Musical Talents might provide an occasion to discuss these issues further.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors report no conflict of interest.

Notes

1. For further reading on the historical distinctions between musicality, musicianship, and musical sophistication, please consult Boyle & Radocy (1987), Gembris (1997), McPherson & Hallam (2016), and Müllensiefen et al. (2014), who have written previously on the difference between terms.
2. The exact detailing of the tests and their creations can be found in the original Gold-MSI paper (Müllensiefen et al., 2014) under the *Study 4: Self-Reported Musical Sophistication and Objective Listening Tests in a Large Sample* subheading.
3. The scripts and data used for the reported analyses can be accessed via the supplemental material on this project's Open Science Framework project page, maintained by the last author.
4. The positive manifold refers to the phenomenon in cognitive psychology that if an individual tends to do well on one cognitive test, such as a test of mental rotation, then they are likely to also do well on a test of general fluid intelligence and working memory capacity (Ritchie, 2015).
5. The Flynn Effect is a well-documented phenomenon in psychology that notes the steady, significant increase in intelligence over time (e.g., Ritchie, 2015).

ORCID iD

David John Baker https://orcid.org/0000-0003-3921-7517

References

Baker, D. J., & Müllensiefen, D. (2017). Perception of leitmotives in Richard Wagner's Der Ring des Nibelungen. *Frontiers in Psychology, 8*(662). doi:10.3389/fpsyg.2017.00662.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York: Routledge.

Bentley, A. (1966). *Musical ability in children and its measurement*. London: Harrap.

Blacking, J. (1974). *How musical is man?* Seattle: University of Washington Press.

Boyle, J. D., & Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York: Schirmer Books.

Chin, T., Coutinho, E., Scherer, K. R., & Rickard, N. S. (2018). MUSEBAQ: A modular tool for music research to assess musicianship, musical capacity, music preferences, and motivations for music use. *Music Perception, 35*, 376–399.

Chin, T., & Rickard, N. S. (2012). The music use (muse) questionnaire: An instrument to measure engagement in music. *Music Perception, 29*, 429–446.

Colley, I. D., Keller, P. E., & Halpern, A. R. (2017). Working memory and auditory imagery predict sensorimotor synchronization with expressively timed music. *The Quarterly Journal of Experimental Psychology, 71*(8), 1–49.

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

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1–12.

Daly, H. R., & Hall, M. D. (2018). Not all musicians are created equal: Statistical concerns regarding the categorization of participants. *Psychomusicology: Music, Mind, and Brain, 28*, 117–126.

Farrugia, N., Jakubowski, K., Cusack, R., & Stewart, L. (2015). Tunes stuck in your brain: The frequency and affective evaluation of involuntary musical imagery correlate with cortical structure. *Consciousness and Cognition, 35*, 66–77.

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae, 17*, 265–276.

Gembris, H. (1997). Historical phases in the definition of musicality. *Psychomusicology: A Journal of Research in Music Cognition, 16*(1–2), 17–25.

Gordon, E. (1989). *Manual for the advanced measures of music audiation*. Chicago, IL: GIA Publications.

Greenberg, D. M., Müllensiefen, D., Lamb, M. E., & Rentfrow, P. J. (2015). Personality predicts musical sophistication. *Journal of Research in Personality, 58*, 154–158.

Harrison, P. M., 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), 3618.

Healy, K. (2017). Fuck nuance. *Sociological Theory*, *35*, 118–127.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124.

Irwing, P., & Hughes, D. J. (2018). Test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 3–48). Hoboken, NJ: Wiley Blackwell.

Jaschke, A. C., Honing, H., & Scherder, E. J. A. (2018). Longitudinal analysis of music education on executive functions in primary school children. *Frontiers in Neuroscience*, *12*(103). doi:10.3389/fnins.2018.00103.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York: Guilford Publications.

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*, 151–177.

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PloS ONE*, *7*(e52508). doi:10.1371/journal.pone.0052508.

Levitin, D. J. (2012). What does it mean to be musical? *Neuron*, *73*, 633–637.

Lin, H., Kopiez, R., Müllensiefen, D., & Wolf, A. (2018, July). *The Chinese version of the Gold-MSI: Adaptation and validation of an inventory for the measurement of musicality in a Taiwanese sample*. Talk presented at the International Conference for Music Perception and Cognition 15, Graz, Austria.

Loehlin, J. C., & Beaujean, A. A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Milton Park, UK: Taylor & Francis.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Lawrence Erlbaum Associates.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491.

McPherson, G. E., & Hallam, S. (2016). Musical potential. In S. Hallam, I. Cross, & M. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 433–448). Oxford, UK: Oxford University Press.

Meinz, E. J., & Hambrick, D. Z. (2010). Deliberate practice is necessary but not sufficient to explain individual differences in piano sight-reading skill: The role of working memory capacity. *Psychological Science*, *21*, 914–919.

Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 116–139.

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*, e89642. doi:10.1371/journal.pone.0089642.

Müllensiefen, D., Harrison, P., Caprini, F., & Fancourt, A. (2015). Investigating the importance of self-theories of intelligence and musicality for students' academic and musical achievement. *Frontiers in Psychology*, *6*(1702). doi:10.3389/fpsyg.2015.01702.

Murphy, C. (1999). How far do tests of musical ability shed light on the nature of musical intelligence? *British Journal of Music Education*, *16*(1), 39–50.

Nichols, B. E., Wöllner, C., & Halpern, A. R. (2018). Score one for jazz: Working memory in jazz and classical musicians. *Psychomusicology: Music, Mind, and Brain*, *28*(2), 101–107.

Ollen, J. E. (2006). A criterion-related validity test of selected indicators of musical sophistication using expert ratings (Unpublished doctoral dissertation). Ohio State University. Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=osu1161705351&disposition=inline

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716.

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition, 100*(1), 1–32.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raven, J., Raven, J., & Court, J. (1998). *Manual for Raven's progressive matrices and vocabulary scales*. Oxford, UK: Oxford Psychologists Press.

Revelle, W. (2014). psych: Procedures for personality and psychological research. Northwestern University, Evanston. R package version 1.1.

Ritchie, S. (2015). *Intelligence: All that matters*. London, UK: Hodder & Stoughton.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Schaal, N. K., Bauer, A.-K. R., & 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 questionnaire for measuring musical sophistication in a German population]. *Musicae Scientiae, 18*, 423–447.

Seashore, C. E. (1915). The measurement of musical talent. *The Musical Quarterly, 1*(1), 129–148.

Seashore, C. E. (1919). *Manual of instructions and interpretations for measures of musical talent*. London, UK: Columbia Graphophone Company.

Seashore, C. E., Lewis, D., & Saetveit, J. G. (1960). *Measures of musical talents [test]*. New York: Psychological Corporation. (Earlier version published 1919, 1939)

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.

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science, 15*, 511–514.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53–61.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Pittsburgh, PA: Psychology Software Inc.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *The American Journal of Psychology, 15*, 201–292.

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2001). *Using multivariate statistics*. Harlow, UK: Pearson Publishing.

Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PLoS ONE, 12*, e0186773.

Tarrant, M., North, A. C., & Hargreaves, D. J. (2002). Youth identity and music. *Musical Identities, 13*, 134–150.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505.

Ventura, J. (2018). Predicting working memory and fluid intelligence from measures of musicality. Unpublished Master's Thesis, Louisiana State University. Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/4755/

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*, 188–196.

Wing, H. D. (1968). *Tests of Musical Ability and Appreciation* (2nd ed.). Cambridge, UK: Cambridge University Press.

Wolf, A., & Kopiez, R. (2018). Development and validation of the Musical Ear Training Assessment (META). *Journal of Research in Music Education, 66*(1), 53–70.