MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

This postprint was originally published by University of California Press as:
Lin, Z., Werner, A., Lindenberger, U., Brandmaier, A. M., & Wenger, E. (2021). **Assessing music expertise: The Berlin Gehoerbildung Scale.** *Music Perception*, *38*(4), 406–421. https://doi.org/10.1525/mp.2021.38.4.406.

Supplementary material to this article is available. For more information see http://hdl.handle.net/21.11116/0000-0007-7670-6

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, nontransferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. By using this particular document, you accept the above-stated conditions of use.

Provided by:

Max Planck Institute for Human Development
Library and Research Information
library@mpib-berlin.mpg.de

This postprint was originally published by University of California Press as:
Lin, Z., Werner, A., Lindenberger, U., Brandmaier, A. M., & Wenger, E. (2021). Assessing music expertise: The Berlin Gehoerbildung Scale. *Music Perception, 38*(4), 406–421. https://doi.org/10.1525/mp.2021.38.4.406.

Assessing Music Expertise: The Berlin Gehoerbildung Scale

Ziyong Lin, Andre Werner,
Ulman Lindenberger,
Andreas M. Brandmaier, & Elisabeth Wenger
Max Planck Institute for Human Development, Berlin, Germany

Correspondence concerning this article should be addressed to Ziyong Lin, Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: ziyong@mpib-berlin.mpg.de.

WE INTRODUCE THE *BERLIN GEHOERBILDUNG* SCALE (BGS), a multidimensional assessment of music expertise in amateur musicians and music Professionals. The BGS is informed by music theory and uses a variety of testing methods in the ear-training tradition, with items covering four different dimensions of music expertise: (1) intervals and scales, (2) dictation, (3) chords and cadences, and (4) complex listening. We validated the test in a sample of amateur musicians, aspiring Professional musicians, and students attending a highly competitive music conservatory ($n = 59$). Using structural equation modeling, we compared two factor models: a unidimensional model postulating a single factor of music expertise; and a hierarchical model, according to which four first-order subscale factors load on a second- order factor of general music expertise. The hierarchical model showed better fit to the data than the unidimensional model, indicating that the four subscales capture reliable variance above and beyond the general factor of music expertise. There were reliable group differences on both the second-order general factor and the four subscales, with music students outperforming aspiring Professionals and amateur musicians. We conclude that the BGS is an adequate measurement Instrument for assessing individual differences in music expertise, especially at high levels of expertise.

Key words: music expertise, musicians, multidimensional assessment, structural equation modeling, hierarchical factor model

Supplementary material to this article is available. For more information see http://hdl.handle.net/21.11116/0000-0007-7670-6.

MUSIC IS AN ESSENTIAL HUMAN EXPERIENCE and exists in every known human culture (Blacking & Netti, 1995). In Western societies, musicians who receive formal music education undergo ear training and training in music theory. Recent years have seen an increasing interest in scientific investigations of music ability. Expert musicians, who often specialize in certain areas and can be ascribed specific roles such as singers, conductors, and instrumentalists, have devoted extraordinary amounts of time to musical practice and thus provide a model for the effects of long-term training on brain and behavior (Herholz & Zatorre, 2012; Schlaug, 2001). Moreover, investigating individual differences in music expertise can help to elucidate the interplay between genetic predisposition and experience (Mosing, Madison, Pedersen, Kuja-Halkola, & Ullén, 2014; Niarchou et al., 2019; Ullén, Hambrick, & Mosing, 2016).

Given the increasing interest in antecedents of individual differences in music expertise, the valid and reliable assessment of individuals' music expertise and skills has gained in importance. In practice, however, studies typically use a binary Classification of musicians versus nonmusicians, which often is based on self-report data, such as years of music practice, music degrees, or both (e.g., Wong, Skoe, Russo, Dees, & Kraus, 2007). While such a classificatory approach might facilitate recruitment and group comparisons, it is of limited utility for process-based analyses of individual differences. First, it does not address individual differences within groups, which might be substantial and serve as a point of departure for delineating behavioral, neural, and genetic correlates of music ability. Second, using a cutoff based on years of playing an instrument or similar experiential measures as proxies of music expertise presupposes that music ability is unidimensional, and that years of deliberate practice convey the level of music expertise that has been attained. Available evidence contradicts both of these Claims, as it indicates that music expertise is a multidimensional construct (Chin, Coutinho, Scherer, & Rickard, 2018; Hallam, 2010), and that the effects of deliberate practice are contingent upon genetic predisposition (Mosing et al., 2014). Third, a critical review of available behavioral and neuroimaging studies by Daly and Hall (2018) has shown that the results of these studies tend to vary as a function of the kind of cutoff that is used for Classification. In contrast, results tended to be more consistent and associated with larger effect sizes when years of training was treated as a continuous variable.

Apart from the typical practice in research studies, there is little agreement on how objective tests can best

measure music ability. The operational definition of music ability appears to differ across different measures. The basic distinction in the literature is between musical aptitude and musical achievement. While the former refers to potential levels of music ability before formal training and education, the latter refers to attained levels of music abilities (Zentner & Gingras, 2019). It needs to be kept in mind, however, that music expertise appears to emanate from the interplay of genetic endowment and environmental factors (Mosing et al., 2014; Ullén et al., 2016); hence, music expertise and music aptitude are not orthogonal but interrelated. In the following, we review already existing tests that assess music ability in one or the other way.

Probably the first documented tool, the *Measure of Musical Talent*, was published 100 years ago (Seashore, 1919). Since then, many instruments and measures have been developed. Among them, aptitude tests that aim to measure future success in music are particularly noteworthy (for review, also see Boyle & Radocy, 1987). Typical examples are the aforementioned *Seashore Measure of Musical Talent* (Seashore, 1919; Seashore, Lewis, & Saetveit,1956), the *Wing Standardized Tests of Musical Intelligence* (Wing, 1948), and the *Gordon Music Aptitude Profile* (Gordon, 1967). Most of these tests have been designed for educational purposes and classroom administration, rather than for assessing differences among adults with music training. Despite the efforts that went into the design of these tests, their actual use in research and practice is limited (Carson, 1998). Law and Zentner (2012) noted some of the possible reasons, including inaccessibility of testing materials, outdated versions and sound files, an exclusive focus on children as well as unknown or inadequate validity and reliability.

More recently, a number of scales and questionnaires have been developed to target adults, including special populations, and to broaden the range of musical skills that are being assessed. For example, the *Montreal Battery of Evaluation of Amusia* (Peretz, Champod, & Hyde, 2003) was designed to assess impairments in pitch, rhythm, and musical memory in order to identify amusia. Several other tests and batteries have been designed for populations with hearing loss and implants (e.g., Kang et al., 2009; Kirchberger & Russo, 2015; Spitzer et al. 2008; Uys & van Dijk, 2011). With respect to questionnaires and batteries targeting adults with normal hearing, a new trend towards self-report instruments has emerged. Most questionnaires are designed to measure a specific construct, such as musical engagement and processing (*Brief Experiences with Music Questionnaire*, Werner, Swope, & Heide, 2006; *Music Empathizing-Systemizing*, Kreutz, Schubert, & Mitchell, 2008; *EMuJoy*, Nagel, Kopiez, Grewe, & Altenmüller, 2007; *Music in Mood Regulation*, Saarikallio, 2008; *Geneva Emotional Music Scale*, Zentner, Grandjean, & Scherer, 2008; *Barcelona Music Reward Questionnaire*, Mas-Herrero, Marco-Pallares, Lorenzo-Seva, Zatorre, & Rodriguez-Fornells, 2013; *Healthy-Unhealthy Music Scale*, Saarikallio, Gold, & McFerran, 2015; *Music Engagement Questionnaire*, Vanstone, Wolf, Poon, & Cuddy, 2016), use of music (*Use of Music Inventory*, Chamorro-Premuzic & Furnham, 2007; *Music Use Inventory*, Lonsdale & North, 2011, *Music in Everyday Life*, Gottfried, Thompson, Elefant, & Gold, 2018), music skills and training (*Ollen Musical Sophistication Index*, Ollen, 2006; *Music Training Questionnaire*, Brod & Opitz, 2012; *Music USE Questionnaire*, Chin & Rickard, 2012), and musical preferences (*Short Test of Musical Preferences*, Rentfrow & Gosling, 2003; *MUSIC Model*, Rentfrow, Goldberg, & Levitin, 2011). Researchers have also developed instruments that aim at capturing the multifaceted nature of music ability and music experience. A self-report instrument of this type is the *Music Use and Background Questionnaire* (MUSEBAQ, Chin, Coutinho, Scherer, & Rickard, 2018), including four subscales to measure musicianship, musical capacity, music preferences, and motivation for music use. Another example, the *Goldsmiths Musical Sophistication Index* (Gold-MSI; Müllensiefen et al., 2014), is composed of a self-report inventory with five sub-scales assessing singing abilities, perceptual abilities, music training, active engagement, and emotion. In addition to the self-report questionnaire, the Gold-MSI includes two objective listening tasks, namely a beat perception task and a melodic memory task (Müllensiefen et al., 2014).

Furthermore, several other new tools have been developed that focus on the quantification of musicianship, such as the *Music Ear Test* (MET, Wallentin et al., 2010), the *Profile of Music Perception Skills* (PROMS; Law & Zentner, 2012), the *Swedish Musical Discrimination Test* (SMDT, Ullén, Mosing, Holm, Eriksson, & Madison, 2014), and the *Musical Ear Training Assessment* (META, Wolf & Kopiez, 2018). Most of these new instruments follow the older traditional aptitude tests, using a ‘‘same’’ versus ‘‘different’’ judgment task for pairs of newly created melodic or rhythmic lines with increasing length and complexity. The experimental stimuli in these tools are often artificially created and use a single instrument (e.g., piano). Most of them give no or little room to multi-instrument textures and sound quality, even though timbre is an important musical dimension (Müllensiefen et al., 2014). Using relatively simple, artificially created stimuli is beneficial for controlling familiarity effects but comes at the expense of ecological validity; that is, these

Originally published in: Music Perception, 38(4), 2021, p. 407

tests often do not reflect the broad range of musical expertise and listening experience of musicians. With years of training and practicing, musicians develop expertise and mastery in their specialized area (e.g., instruments, voice, composition). They also learn to attend to and process fine-grained acoustic and aesthetic properties of musical sounds. Through formal music training, musicians acquire musical notation, concepts, and terminology to dictate, analyze, and create musical material, as well as communicate about this material with other musicians. These nuances at the high end of the musical expertise spectrum are not fully captured by currently available tools.

Here, we present a new test of music expertise, the *Berlin Gehoerbildung Scale* (BGS). Gehoerbildung is the German term for ear training. The BGS aims at assessing various aspects of music expertise in the tradition of Western art music and is intended as a complementary test to already existent measures of music expertise. We would like to emphasize that this concept of music expertise here is the interactive result of musical talent and practice, does require formal music education and training and is situated at the upper end of music achievement. For instance, musicians can display great musicality in playing an instrument without having attained a high level of formal music education. To achieve a focus on formal music training, the BGS is structured into four subscales, each composed of several complex items, which together provide a fairly wide coverage of musical knowledge and abilities of formally trained musicians and music professionals. The subscales are named *Intervals and Scales*, *Dictation*, *Chords and Cadences*, and *Complex Listening*. The *Intervals and Scales* subscale assesses one's ability to name and notate musical intervals and scales. The subscale *Dictation* includes items of melodic and rhythmic dictation, which requires one to notate the presented rhythms, melody, voices, or all together. The subscale *Chords and Cadences* requires one to recognize, name, and notate chords, cadences, and chord progressions, as well as to understand harmonic functions. Finally, the subscale *Complex Listening* taps into more complex listening skills, such as identifying scoring deviations in a given musical piece, and recognizing instruments in a musical excerpt.

To overcome some of the limitations of already available measures, the BGS uses a variety of testing methods other than same-different judgments. Individuals are required to use musical terminology and musical notations during the assessment. Moreover, rather than using artificial stimuli, the BGS uses music excerpts from the Western art music tradition to ensure ecological validity. Musical samples are from high-quality recordings and are selected from a range of timbres from multiple instruments. More importantly, in addition to more widely assessed basic dimensions such as intervals, scales, and harmony, the BGS also includes items pertaining to complex musicrelated skills, such as the ability to identify scoring deviations in a given musical piece. To validate and analyze the psychometric properties of our scale, we tested three groups that are likely to differ on the dimensions assessed by the BGS: (a) amateur musicians who are actively engaged in music activities; (b) aspiring professional musicians who are preparing for their entrance exam at a highly competitive music conservatory; (c) music students who are currently undergoing formal training in classical Western music at a music conservatory.

Our study was informed by three general expectations that together characterize the BGS as a measure of music expertise. First, we predict that the BGS would measure music expertise in a continuous fashion by items with satisfactory inter-rater agreement. Second, we expect that its psychometric structure will be well captured by a general dimension of *Music Expertise*, and four specific subscales (*Intervals and Scales*, *Dictation*, *Chords and Cadences*, and *Complex Listening*). Third, orthogonal group comparisons will reveal significant group differences between (i) the amateur musicians and the two groups with additional formal training; and (ii) the group of aspiring professional musicians and the group of students who are already enrolled at a music conservatory.

Method

PARTICIPANTS

A total of 59 individuals aged 17 to 33 years ($M_{age} = 22.25$, $SD = 3.81$, 33.9% female) were recruited through flyers, mailing lists, project presentations in music schools, and word-of-mouth recommendation in Berlin, Germany. Individuals belonged to one of three groups, amateur musicians ($n = 17$) who were actively performing music in everyday life, aspiring professional musicians ($n = 23$) who were in the process of preparing for the entrance exam for a music conservatory, and music students ($n = 19$) who were currently enrolled at a music conservatory. All participants either sang or played at least one primary instrument, and had at least five or more years of experience singing or playing their instrument. Years of singing or playing a primary instrument ($M_{year} = 12.54$, $SD = 5.05$) were comparable across the three groups, $F(2, 57) = 0.211$, $p = .81$ (amateur musicians: $M_{year} = 12.74$, $SD = 5.97$; aspiring professional musicians: $M_{year} = 12.00$, $SD = 4.67$; music students:

M_{year} = 13.00, SD = 4.81; one participant in the aspiring professional group did not provide information about his or her primary instrument or years of playing). In terms of education, 8.5% of the participants had completed junior high school (Mittlere Reife), 66.1% of them had completed high school (Abitur), and 25.4% of them had already attained a college degree or equivalent education (Abgeschlossenes Studium). When comparing across groups, participants in three groups were similar in age, $F(2, 58) = 0.620$, p =.54 (amateur musicians; M = 23.12, SD = 3.43; aspiring professional musicians: M = 22.00, SD = 3.78; music students: M = 21.79, SD = 4.24) and years of education, F (2, 58) = 2.98, p =.06 (amateur musicians: M = 14.32, SD = 3.77; aspiring professional musicians: M = 12.30, SD = 3.11; music students: M = 12.34, SD =.94). All participants had normal hearing, did not have any metallic implants, and had not had any psychiatric diagnosis. Participants were part of a larger study and additionally underwent comprehensive structural and functional magnetic resonance imaging (MRI). Participants were paid up to 200€ for completion of the whole study (including up to 5 measurement time points with 1.5 h of MRI and 1.5 h of behavioral testing). The ethical board of the DGPs (Ethikkommission der Deutschen Gesellschaft für Psychologie) approved the study, and written consent of all participants (including informed consent by the guardians of one 17-year old participant) was obtained prior to investigation.

MATERIALS

The *Berlin Gehoerbildung Scale* (BGS; see Supplementary Material accompanying the online version of this paper at http://hdl.handle.net/21.11116/0000-0007-7670-6 for the complete task, audio samples, answer key, and instructions) was designed by André Werner, a composer, collaborator, and co-author of this study. The BGS requires listening to musical recordings, and the use of musical notation. It taps into various aspects of knowledge and skill in ear training and music theory, including intervals, scales, dictation, rhythm, harmony, identifying deviations in music excerpts, and instrument recognition (see below). The BGS consists of 12 items; see Table 1 for item names, their abbreviations, and maximum attainable scores. (Originally, the BGS also contained another item, called Identifying Meter. This item did not correlate with any of the other items and was removed from further analysis.) Item 1 (Naming Intervals) requires participants to listen to 10 intervals once, and one at a time. They are asked to provide the name and direction (ascending, parallel, or descending) of each interval. Item 2 (Notating Intervals) provides the initial or lower tone for each interval. Participants listen to 10 intervals once, and one at a time, and complete the intervals by providing the second/higher notes. In Item 3 (Naming Scales) participants listen to three scales twice and are required to provide the name and direction of each scale. In Item 4 (Naming and Notating Scales) they again listen to three scales twice and are asked to write down each scale using music notations. For Item 5 (Melodic Dictation) participants listen to three short melodies three or four times (first two melodies for three times, last melody for four times) and write them down. Item 6 is Rhythmic Dictation, which includes two short music excerpts. Participants listen to each of them twice and they must notate the rhythm as it is heard. Item 7 tests participants' ability to name chords. Eight chords are played once, and one at a time. Participants are asked to provide the name and position of each chord. Item 8, Naming and Notating Chords, requires participants to name and notate each of the five chords as heard. Item 9 focuses on cadences and functions. Participants listen to two music excerpts (each for three times) and write down the chords and cadences as well as their functions. Item 10 and Item 11 both require participants to listen to a music excerpt, compare the given score to the sample they hear, and annotate deviations or errors (i.e., incorrect notes, extra measures) in the score. While the excerpt of Item 10 is from a piano piece, the excerpt of Item 11 is from a string quartet. The last item, Recognizing Instruments, includes two pieces of music. Participants are asked to identify the instruments that can be heard in each piece.

To reach a wider audience in the research community, we have translated the BGS into English, and provided both the English and the original German versions in the Supplementary Material (see http://hdl.handle.net/21.11116/0000-0007-7670-6). An American researcher, who speaks fluent German and was not part of the study team, independently evaluated the translation. All authors are highly proficient in English, and all except the first author speak German as their native language. We checked for effective communication of the original meaning in a manner that is both readable and comprehensible. Given that the BGS is based on formal music theory, two of the authors, who hold music degrees, specifically checked accuracy, consistency, and proper conventions of music terminology in the test. For example, in Item 9 Cadence and Functions, we included solutions using both German functional theory (taught in Germany) and Roman numeral analysis (taught in other countries, such as the United States). The objective nature of the assessment along with our translation effort is likely to reduce the noise introduced by the translation. Having said that, we note that the English version of the BGS still awaits future validation in an independent sample.

Originally published in: Music Perception, 38(4), 2021, p. 409

TABLE 1. *Items and Maximum Attainable Scores in the Berlin Gehoerbildung Scale*

	Abbreviation	Maximum
1. Naming Intervals	NaInt	20
2. Notating Intervals	NotInt	20
3. Naming Scales	NaSca	18
4. Naming and Notating Scales	NotSca	18
5. Melodic Dictation	MelDic	178
6. Rhythmic Dictation	RhyDic	90
7. Naming Chords	NaCh	24
8. Naming and Notating Chords	NotCh	25
9. Cadence and Functions	Cad	81
10. Identifying Deviations (Piano)	DevP	80
11. Identifying Deviations (String Quartet)	DevS	81
12. Recognizing Instruments	RecIns	128
Total		763

PROCEDURE

All of the participants completed the BGS at the Max Planck Institute for Human Development in Berlin, Germany. Testing was untimed and took 60 to 90 minutes, depending on the speed with which participants worked on the various items. The test taking time also includes the length of pauses participants took between tasks and time to respond. The number of repetitions for each sample was preset and the same across participants. Participants were individually tested in a quiet room. For each item, participants received a short instruction on how often they will be presented with a specific stimulus and were asked to complete the items on an answer sheet; for details, please see the SM. Audio files were played on an iMac using the VLC player (version 2.0.0, developed by VideoLAN) and two JBL Control 2 Pro loudspeakers, with the volume of the audio adjusted to a comfortable level. No feedback was given while participants completed the test.

ANALYSIS PLAN

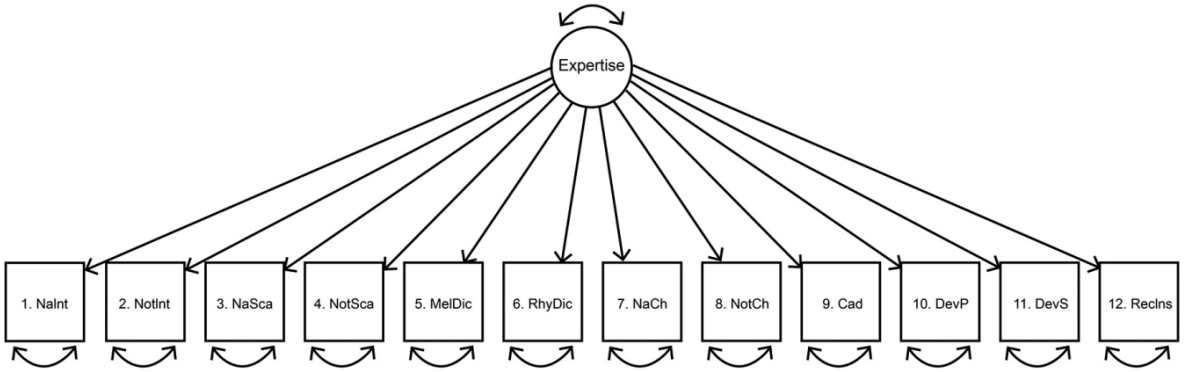
Reliability. Two raters scored a random subset of 40 participants' responses, based on an answer key that had been prepared beforehand. The two raters were professional musicians and experts in music theory; one of them was the test designer. Both were blind to which group the participants were in. To assess inter-rater agreement, their scorings were compared by means of Intraclass Correlation Coefficients (ICC2) based on SPSS version 24, with a two-way mixed effect model and absolute agreement definition.

Suitability for Factor Analysis. Before setting up the models, we tested data suitability for factor analysis. The correlation matrix was computed to inspect if the correlation coefficients were over.30 (Williams, Onsman, & Brown, 2010). Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (should be above .50) and Bartlett's Test of Sphericity (should be significant) were also conducted (Williams et al., 2010).

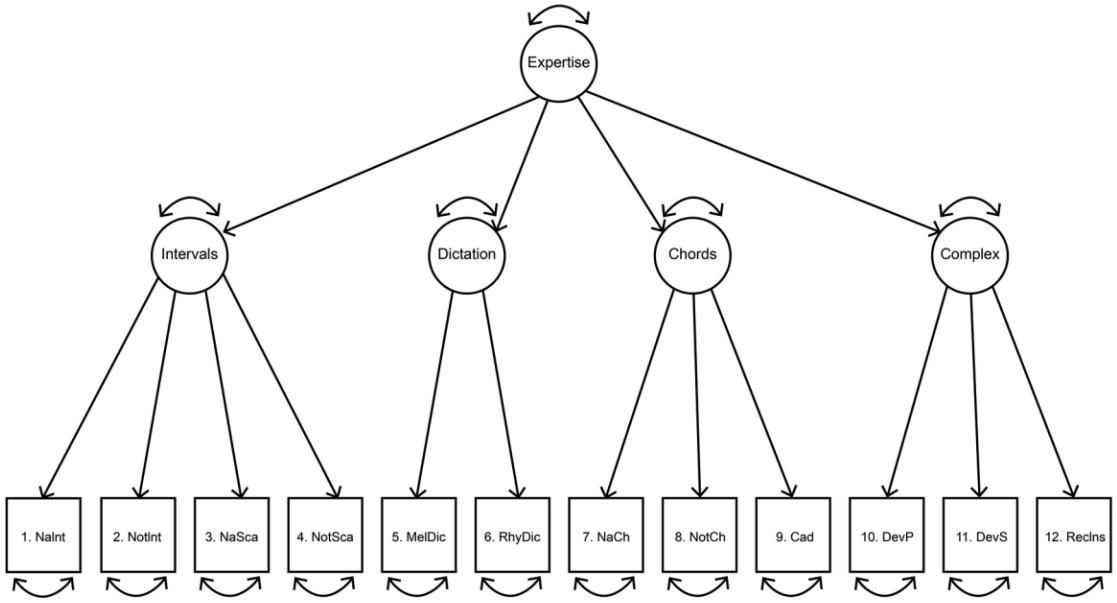
Model Specification and Comparison. We expected that the BGS would form four subscales that together define a general factor of *Music Expertise*. Specifically, we assumed that Items 1 to 4 would form the subscale *Intervals and Scales*, given that these items focused on naming and notating musical scales and intervals. Items 5 and 6 consisted of melodic and rhythmic dictations, and were expected to define the subscale *Dictation*. In Items 7 to 9, participants had to name and notate chords and cadences as well as their functions; therefore, we expected these to form the subscale *Chords and Cadences*. Finally, in Items 10 to 12, participants had to engage in various complex listening skills, such as indicating deviations between a given music score and the musical piece which was played, resulting in a putative subscale called *Complex Listening*.

To determine the dimensionality of the BGS, we fitted two different models to the data (see Figure 1). Model 1 postulated a single, first-order factor of music expertise. This model assumes that all items of the BGS form a single general ability factor, which means that there is no shared variance between the items of any of the four subscales over and above the common factor. The alternative model is a hierarchical model allowing for subscale-specific variance between persons. In this model, the four subscales are each represented by a first-order factor, and together they load onto a second-order general factor of music expertise. The hierarchical model assumes that the higher-order general factor of expertise accounts for the commonality among the four subscales. We also fitted a bifactor model to the

Originally published in: Music Perception, 38(4), 2021, p. 410



Model 1: Unidmensional Model



Model 2: Hierarchical Model

FIGURE 1. Illustration of the two proposed models of music expertise. Single headed arrows are regressions, double-headed arrows are variances. Circles represent latent variables and rectangles represent observed variables namely items 1 to 12. Expertise: Music Expertise; Intervals: Intervals and Scales; Chords: Chords and Cadences; Complex: Complex Listening. For items, 1.NatInt: Naming Intervals; 2, NotInt: Notating Intervals; 3.NaSca: Naming Scales; 4.NotSca: Naming and Notating Scales; 5.MelDic: Melodic Dictation; 6.RhyDic: Rhythmic Dictation; 7.NaCh: Naming Chords; 8.NotCh: Naming and Notating Chords; 9.Cad: Cadence and Functions; 10.DevP: Identifying Deviations (Piano); 11.DevS: Identifying Deviations (String Quartet); 12.RecIns: Recognizing Instruments.

data. The bifactor model showed almost identical fit to the hierarchical model.[1]

Formal comparisons between the two models were used to assess the structure of music expertise as measured by the BGS. If a single general factor of music

[1] The factor score estimates of music expertise between the bifactor model and hierarchical model were strongly correlated, $r(57) = .98$, $p < .0001$, so we decided to not report it in the main text. For details, please see Supplementary Materials at paper at http://hdl.handle.net/21.11116/0000-0007-7670-6.

Originally published in: Music Perception, 38(4), 2021, p. 411

TABLE 2. *Descriptive Summary of Item Scores Across Groups*

Items	Amateur Musicians ($n = 17$) M (SD)	Aspiring Professionals ($n = 23$) M (SD)	Music Students ($n = 19$) M (SD)	Total Sample ($n = 59$) M (SD)
1. NaInt	6.68 (4.27)	13.58 (4.21)	17.05 (3.39)	12.73 (5.68)
2. NotInt	6.50 (4.21)	11.94 (3.75)	16.71 (3.79)	11.91 (5.52)
3. NaSca	7.00 (4.70)	13.38 (4.01)	16.37 (2.48)	12.52 (5.29)
4. NotSca	3.06 (3.86)	13.25 (5.64)	14.74 (3.90)	10.83 (6.77)
5. MelDic	12.18 (14.78)	30.91 (27.53)	105.32 (54.55)	49.47 (53.27)
6. RhyDic	26.53 (17.24)	40.07 (13.09)	69.42 (13.78)	45.62 (22.60)
7. NaCh	3.41 (3.14)	10.00 (7.83)	18.63 (5.28)	10.87 (8.26)
8. NotCh	.53 (1.23)	9.29 (7.94)	16.79 (7.50)	9.18 (9.07)
9. Cad	2.47 (6.85)	22.52 (24.36)	53.00 (25.27)	26.56 (29.01)
10. DevP	7.53 (10.38)	17.04 (17.11)	41.26 (22.51)	22.10 (22.15)
11. DevS	5.88 (7.12)	10.39 (12.47)	35.21 (30.19)	17.08 (22.76)
12. RecIns	62.12 (25.79)	79.17 (15.76)	91.26 (19.49)	78.17 (22.87)

TABLE 3. *Floor and Ceiling Effects Across Groups*

Items		Amateur Musicians ($n = 17$) Frequency (Percentage)	Aspiring Professionals ($n = 23$) Frequency (Percentage)	Music Students ($n = 19$) Frequency (Percentage)
1. NaInt	Floor	1(6%)	-	-
	Ceiling	-	2(9%)	6(32%)*
2. NotInt	Floor	3(18%)*	-	-
	Ceiling	-	-	3(16%)*
3. NaSca	Floor	2(12%)	-	-
	Ceiling	1(6%)	7(30%)*	13(68%)*
4. NotSca	Floor	9(53%)*	-	-
	Ceiling	-	9(39%)*	6(32%)*
5. MelDic	Floor	5(30%)*	1(4%)	-
	Ceiling	-	-	3(16%)*
6. RhyDic	Floor	2(12%)	-	-
	Ceiling	-	-	2(11%)
7. NaCh	Floor	5(30%)*	2(9%)	-
	Ceiling	-	3(13%)	5(26%)*
8. NotCh	Floor	14(82%)*	2(9%)	-
	Ceiling	-	1(4%)	3(16%)*
9. Cad	Floor	14(82%)*	3(13%)	-
	Ceiling	-	-	1(5%)
10. DevP	Floor	9(53%)*	7(30%)*	1(5%)
	Ceiling	-	-	1(5%)
11. DevS	Floor	9(53%)*	9(39%)*	3(16%)*
	Ceiling	-	-	4(21%)
12. RecIns	Floor	1(6%)	-	-
	Ceiling	-	-	-

Note: * indicates frequency > 15%; - indicates that no participant scored at the lowest or highest level of the measurement range.

expertise would fit just as well or better as the hierarchical models, then there would be no need to posit the existence of the four subscales. Instead, music expertise would be well described by one general dimension. On the other hand, a comparatively good fit of the hierarchical model would suggest that the structure of music expertise resembles the positive manifold of intellectual abilities, where broad cognitive abilities are seen as expressions of general intelligence (e.g., Carroll, 1993).

Originally published in: Music Perception, 38(4), 2021, p. 412

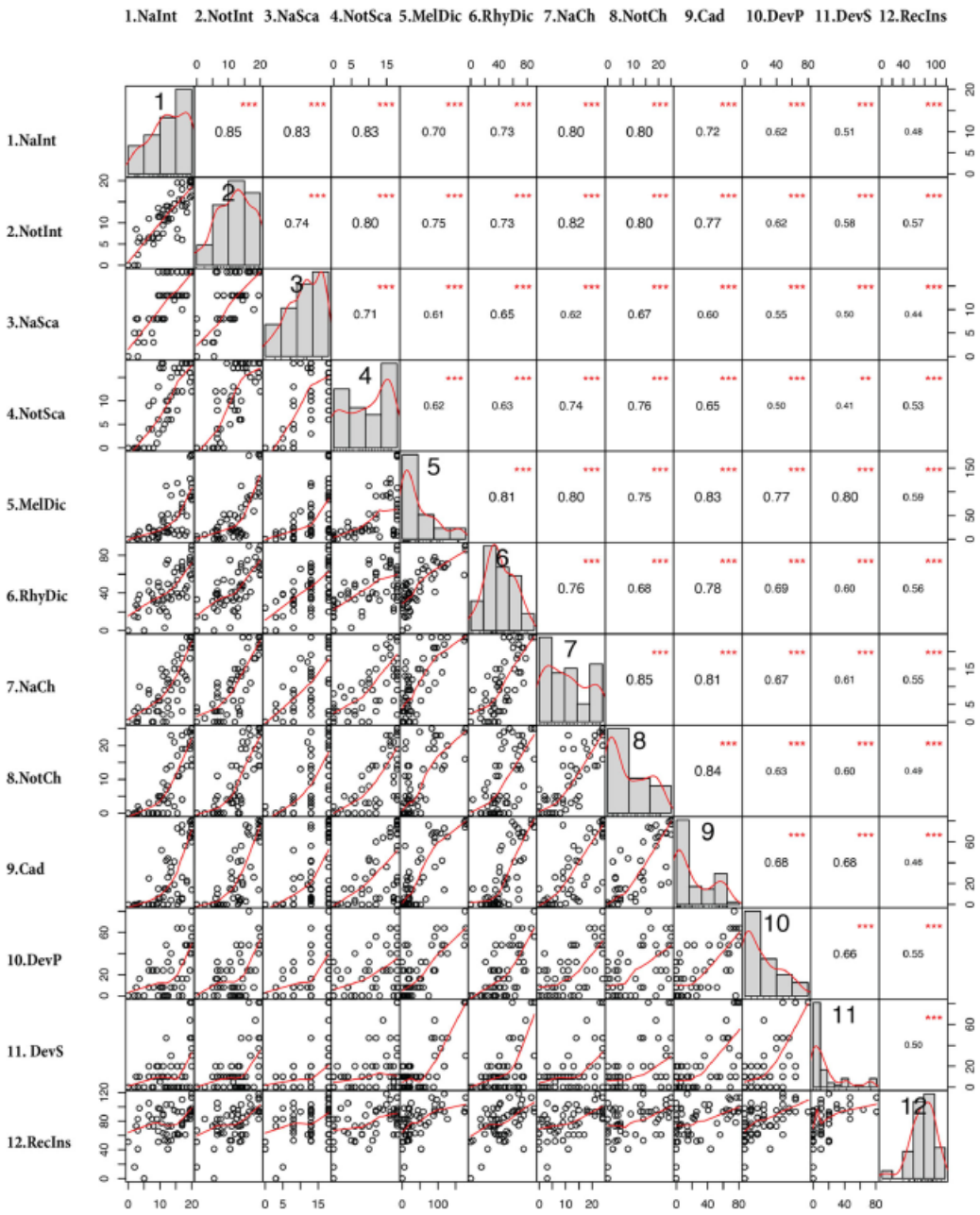


FIGURE 2. Item score distributions and correlations. Axes represent test scores, dots are individual data points, and the main diagonal histograms represent frequency counts for each item. Numbers in the main diagonal refer to the 12 items of the BGS.

The two models were specified and estimated in R 3.6.1 using lavaan (Rosseel, 2012) and using the Structural Equation Modeling software Ωnyx (von Oertzen, Brandmaier, & Tsang, 2015) for graphical display. The two models are also provided as Ωnyx xml files in SM. For each model, we report degrees of freedom and chi-square values as well as the Comparative Fit Index (CFI), the standardized root mean square residual (SRMR), and the root square error of approximation (RMSEA) as indices of model fit. According to convention (Hooper, Coughlan, & Mullen, 2008), models with a CFI of .95 or higher, a SRMR of .08 or lower, and a RMSEA of .08 or lower are deemed acceptable. In addition to inspecting the fit of each model individually,

TABLE 4. *Intraclass Correlation Coefficient Estimates and 95% Confident Intervals Between Two Raters*

Items	Intraclass Correlation ($n = 41$)	95% Confidence Interval	
		Lower Bound	Upper Bound
1. NaInt	.995	.991	.997
2. NotInt	.979	.932	.991
3. NaSca	1.00	-	-
4. NotSca	.949	.896	.974
5. MelDic	.981	.946	.992
6. RhyDic	.943	.896	.970
7. NaCh	.964	.932	.980
8. NotCh	.979	.960	.989
9. Cad	.966	.937	.982
10. DevP	.942	.893	.969
11. DevS	.889	.800	.939
12. RecIns	.968	.937	.983

we tested whether the hierarchical model provided significantly better fit to the data than the unidimensional model using a likelihood ratio test.

Testing Group Differences in Music Expertise at the Latent Level. After accepting the better model of BGS, we tested group differences between musicians. To do so, we used dummy variables and we set up two orthogonal contrasts to test for group differences in the latent variables: the first putting amateur musicians against aspiring professional musicians and music conservatory students, and the second putting aspiring professional musicians against music conservatory students.

Results

DESCRIPTIVE STATISTICS

Group means and other descriptive statistics for each item of the BGS are shown in Table 2. Figure 2 visualizes item score distributions and pair-wise correlations. As shown in Table 3, some of the items showed floor or ceiling effects for some of the groups in the sense that 15% or more of the individuals in a given group performed at the lowest or highest end of the scale on a given item, respectively (Lim et al., 2015).

RELIABILITY

The scorings of two raters showed very good inter-rater consistencies, with ICC2 equal to or greater than .889 (see Table 4 for exact ICC estimates of all items).

SUITABILITY FOR FACTOR ANALYSIS

Before setting up the models, we tested data suitability for factor analysis. The correlation matrix showed that all variables had at least one correlation coefficient greater than .40 (Figure 2). The Kaiser-Meyer-Olkin measure of sampling adequacy was .926, which is above the recommended value of .5. Bartlett's test of sphericity was significant, χ^2 (78) = 725.25, $p < .001$, suggesting factorability of the correlation matrix. Finally, communalities were above .40, indicating that each factor item shared substantial variance with other items.

UNIDIMENSIONAL MODEL

The first model was a unidimensional model of general music expertise in which all items load onto a single general factor. This model and its standardized factor loadings are presented in Figure 3. Proportion of variance explained (R^2) for each item by the latent structure is shown in Table 5. The unidimensional model did not achieve acceptable fit of the data $\chi^2 = 140.75$, $df = 54$, CFI = .88, SRMR = .06, RMSEA = .17, 90% CI [.13, .20].

HIERARCHICAL MODEL

In the hierarchical model, the subscale factors load onto a second-order factor of general music expertise. The hierarchical model and its standardized factor loadings are presented in Figure 4. Variance Explained (R^2) for each item by the latent structure is shown in Table 5. This model had adequate fit, even though the RMSEA was slightly above conventional thresholds, $\chi^2 = 83.53$, $df = 50$, CFI = .95, SRMR = .05, RMSEA = .11, 90% CI [.06, .15].

MODEL COMPARISONS

The unidimensional model did not show adequate fit, whereas the fit of the hierarchical model was adequate. A nested comparison confirmed that the hierarchical model fitted the data better than the unidimensional model, $df = 4$, $\chi^2 = 57.22$, $p < .01$. This together with the non-adequate fit indices of the unidimensional model makes the hierarchical model our preferred choice. We accepted the hierarchical model as the better representation of the BGS.

TESTING GROUP DIFFERENCES IN MUSIC EXPERTISE AT THE LATENT LEVEL

Using the hierarchical model, we tested whether the three groups of musicians in the present study differed on the general factor of music expertise and, additionally, whether there were any differences in the four

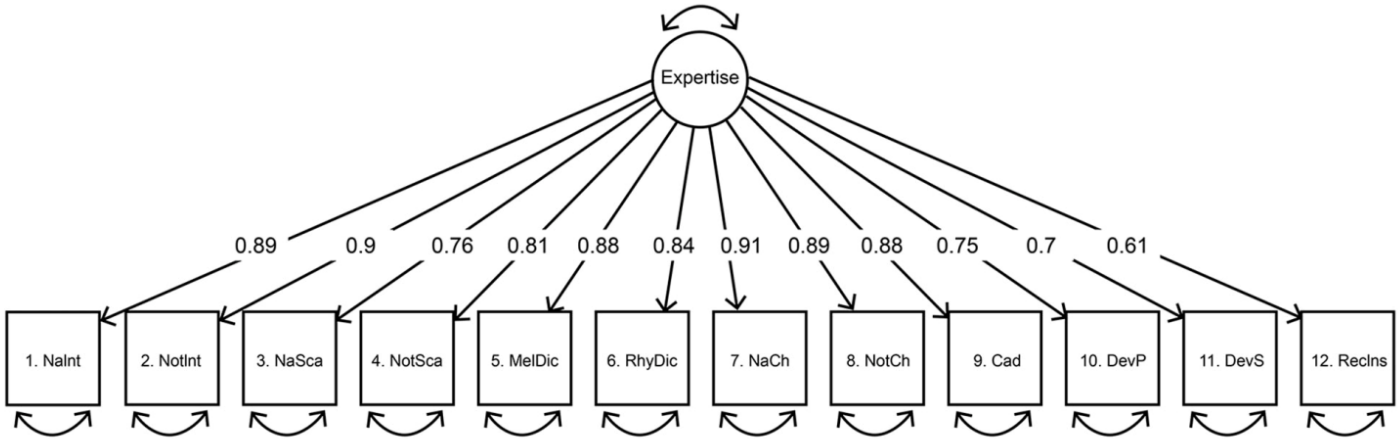


FIGURE 3. Unidimensional model with standardized parameter estimates.

subscales over and above those explained by the general factor. Table 6 reports the effect sizes of the group contrasts for each of the five factors expressed as *variance explained* by group membership. All ten group comparisons were statistically significant and effect sizes ranged from 7% to 56%. To visualize the group differences at the latent subscales, we estimated factor scores in the hierarchical model for all five factors. The resulting plots are shown in Figure 5. As expected, music students scored the highest, aspiring professionals followed, and amateur musicians scored the lowest both on the general factor and the subscale factors.

Next, we tested which of the first-order group differences would remain statistically significant after controlling for group differences in the general factor. Table 7 reports the variance explained by group membership in each of the four subscale factors when controlling for group differences in the general factor of music expertise. First, the group comparison for Intervals and Scales between amateur musicians vs. aspiring professional and music conservatory students was statistically significant. Differences between amateur musicians vs. aspiring professional and music conservatory students favoring the latter two groups remain present after controlling for general music expertise. Several point estimates became negative but none of them was statistically different from zero. Negative estimates may indicate potential suppressor effects in the sense that some specific factors are less sensitive to certain group differences than others. The strongest, but still nonsignificant, evidence for such an effect is seen in Contrast 2 for Intervals and Scales ($z = 1.93$), suggesting that the second-order factor music expertise yields a better separation of professional musicians from music conservatory students than the first-order factor Interval and Scales. Second, the group comparison for Dictation between aspiring professional musicians and music conservatory students favoring the latter also was statistically significant, indicating that the advantage of music conservatory students over aspiring professional musicians was greater than predicted by group differences in general music expertise. The remaining contrasts were not significant. This means that even after controlling for general group differences in the overall factor ‘‘Music Expertise,’’ the subscale factor Intervals and Scales still helped to discriminate between amateur musicians vs. aspiring professionals and music students. Likewise, the subscale factor Dictation still helped to discriminate between aspiring professionals and music conservatory students over and above general group differences as represented by the overall factor ‘‘Music Expertise.‘‘

Discussion

The aim of this present research was to develop a new test for assessing music expertise in trained musicians and music professionals. We have named this tool the *Berlin Gehoerbildung Scale* (BGS). Scale items were generated based on formal music theory to assess music expertise with respect to intervals and scales, dictation, chords and cadences, and complex listening. We recruited and tested amateur musicians, aspiring professional musicians preparing for their entrance exam at a music conservatory, and music students already enrolled at a conservatory. The primary target audience for our BGS are researchers interested in studying professional musicians, while the scale could also be useful in a music education setting to assess students' level of

Originally published in: Music Perception, 38(4), 2021, p. 415

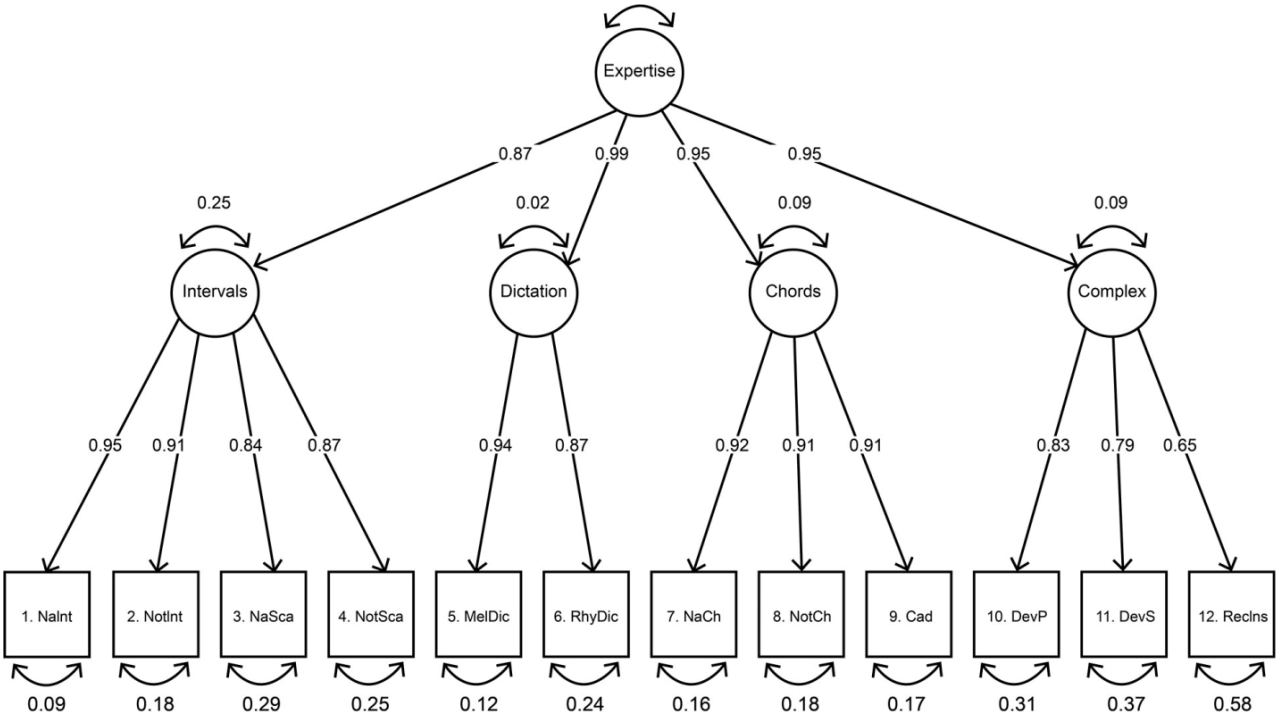


FIGURE 4. Hierarchical model with standardized parameter estimates.

TABLE 5. *Proportion of Variance Explained (R^2) for Each Item Across the 2 Models*

Items	Unidimensional Model	Hierarchical Model
1. NaInt	.79	.91
2. NotInt	.81	.82
3. NaSca	.58	.71
4. NotSca	.66	.75
5. MelDic	.77	.88
6. RhyDic	.71	.76
7. NaCh	.82	.84
8. NotCh	.80	.82
9. Cad	.78	.83
10. DevP	.56	.69
11. DevS	.48	.63
12. RecIns	.37	.42

formal musical education, and ear training skills rather than musical talent.

Inter-rater reliability of the BGS was assessed by two musically trained raters and showed high agreement for all items. To confirm the dimensionality of the BGS and thereby better understand music expertise, we compared two models: a unidimensional model and a hierarchical model. Across models, loadings of most items onto the general factor were strong and generally of similar magnitude, suggesting that the general factor is important in accounting for covariation among items. Across models, item 12 had the lowest standardized loading and, accordingly, the largest amount of variance not explained by the test's factor structure (see Table 5). Item 12 was designed to test participants' ability of processing musical timbre by asking them to name instruments in musical excerpts. Compared to other aspects of musical abilities, the ability to process timbre (e.g., recognizing an instrument) is trained foremost in music students aspiring to become composers, Tonmeister[2] or conductors, and less in students studying to become instrumentalists, or in amateur musicians. This might help to explain why Item 12 showed a rather large amount of specific variance.

A comparison of fit indices between the two tested models showed the hierarchical model to be superior to the model positing a first-order general factor of music expertise. The hierarchical model reflects the assumption that the BGS assesses four musical skills whose

[2] Literally, Tonmeister means sound master. The term describes a person with a profile of music expertise that emphasizes ear training and harmony, who also possesses theoretical and applied knowledge in sound recording, sound production, and audio engineering. University degrees designated as Tonmeister are not restricted to German-speaking countries.

TABLE 6. *Summary of Parameter Estimates for Group Differences*

Factors		R^2	Standardized Estimates	Standard Errors	z	95% CI
Expertise	Contrast 1	.40	.63	.07	8.50	[.49, .78]
	Contrast 2	.27	.52	.08	6.42	[.36, .68]
Intervals	Contrast 1	.56	.75	.06	12.63	[.63, .86]
	Contrast 2	.07	.27	.09	3.18	[.10, .44]
Dictation	Contrast 1	.29	.54	.08	6.80	[.39, .70]
	Contrast 2	.38	.62	.07	8.46	[.48, .77]
Chords	Contrast 1	.42	.65	.07	9.07	[.51, .79]
	Contrast 2	.18	.43	.09	5.10	[.27, .60]
Complex	Contrast 1	.26	.51	.10	5.30	[.32, .70]
	Contrast 2	.28	.53	.10	5.54	[.34, .71]

Expertise: Expert Music Expertise; Intervals: Intervals and Scales; Chords: Chords and Cadences; Complex: Complex Listening; Contrast 1: amateur musicians vs. aspiring professional and music conservatory students; Contrast 2: aspiring professional musicians vs. music conservatory students; R^2 is computed as the square of standardized estimates.

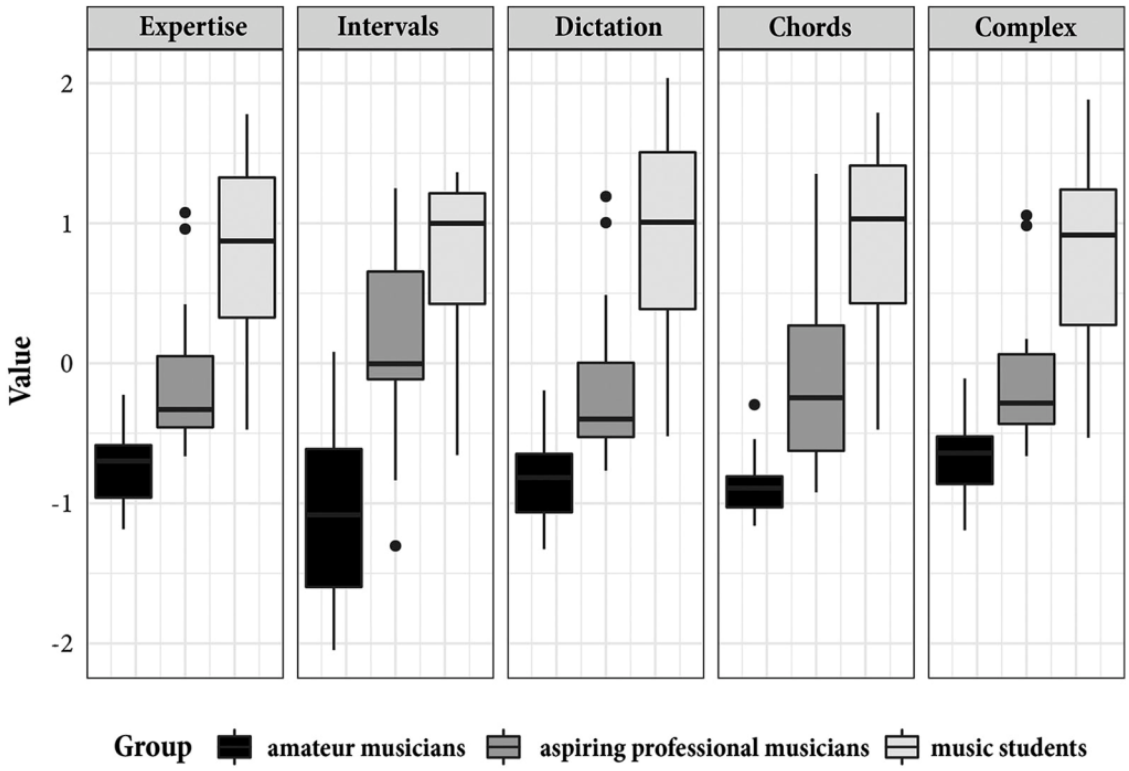


FIGURE 5. Factor score estimates for all five latent factors (the general factor 'Expertise' and the four subscales Intervals, Dictation, Chords, and Complex) across groups of musicians.

commonality is captured by a second-order factor of general music expertise. A strong general ability reflects, for example, that someone who is better at Dictation than the average musician is also likely to perform better at tasks involving Chords and Cadences than the average musician. It is worth noting that loadings on the general factor tended to be higher for Dictation, Chords, and Cadences as well as Complex Listening than for Intervals and Scales, suggesting that the latter is a somewhat weaker indicator of general music expertise. The fact that Intervals and Scales showed a lower loading onto the general factor means that this skill shares less variance with general music expertise, and accordingly has more specific variance than the other skills. In other words, knowing that someone scores high on the factor Intervals and Scales tells us less about

TABLE 7. *Summary of Parameter Estimates for Group Differences in First-order Factors Controlling for the General Factor of Expertise*

Factors		R^2	Standardized Estimates	Standard Errors	z	95% CI
Intervals	Contrast 1	.08	.29	.10	2.78	[.09, .50]
	Contrast 2	.04	-.19	.10	-1.93	[-.38, .00]
Dictation	Contrast 1	.03	-.16	.11	-1.54	[-.37, .04]
	Contrast 2	.03	.18	.08	2.29	[.03, .34]
Chords	Contrast 1	.02	.13	.11	1.13	[-.09, .34]
	Contrast 2	.01	-.09	.09	-.99	[-.27, .09]
Complex	Contrast 1	.04	-.20	.13	-1.49	[-.46, .06]
	Contrast 2	.00	-.03	.13	-.24	[-.28, .22]

Contrast 1: amateur musicians vs. aspiring professional and music conservatory students; Contrast 2: aspiring professional musicians vs. music conservatory students

their general music expertise than the other factors would do. As a note of caution, we acknowledge that the pattern of ceiling and floor effects might also have contributed to this pattern of results.

We hypothesized that the BGS would discriminate among the three groups that were tested. In other words, we expected to be able to tell a specific person's group membership based on their BGS score. Formal group comparisons based on the hierarchical model revealed significant differences between the amateur musicians and the two groups with additional formal training, and between the group of aspiring professional musicians and the group of students who are already enrolled at a music conservatory. As predicted, although the three groups of musicians had comparable years of actively playing music, group differences were statistically significant on both the second-order general music expertise factor as well as on the four subscale factors (see Table 6). In general, music students scored the highest, aspiring professionals followed, and amateur musicians scored the lowest (see Figure 5). Differences between amateur musicians to aspiring professionals on the one side, and between aspiring professionals and music students on the other side, tended to be equally spaced. In follow-up analyses, we found that group differences between amateur musicians vs. aspiring professional and conservatory students remained significant for the Intervals and Scale factor after controlling for the general expertise factor while there was a tendency for a suppression effect between aspiring professional and conservatory students. We conclude that the factor Intervals and Scale differentiates better between amateurs and the other two groups than predicted by general music expertise, but that it differentiates less well between aspiring professionals and conservatory students than predicted by general music expertise. This matches the pattern of group differences shown in Figure 5. From visual inspection, one can see that the gap between the scores of amateur musicians and aspiring professionals appears larger for Intervals and Scales than for the other factors (except for Dictation, see below). At the same time, the scores of aspiring professionals and music students seem to be closer together than on other factors. One might speculate that musicians do acquire some basic knowledge about Intervals and Scales rather early during their education but still are more likely to make mistakes as long as they are amateur musicians. Those reaching aspiring professional levels, get significantly better in this skill and master it sooner than the other musical skills, leaving little to no room for improvement in comparison to professional levels (aka music students). Furthermore, the group differences between aspiring professional musicians and music conservatory students remained significant for the factor Dictation when controlling for the general music expertise factor. That is, we see the pattern (opposite to what we described for Intervals and Scales) that Dictation actually better discriminates between aspiring professionals and music students than the other skill factors, and the general factor. This again is corroborated by Figure 5, where we can see that the gap between the scores of aspiring professionals and music students appears larger in the factor Dictation than in other factors. One possible explanation is that Dictation is a challenging skill in ear training. Although aspiring professional musicians might practice dictation to prepare for attending a music conservatory, it is only at the music conservatory that music students receive rigorous and consistent training in dictation. Likewise, dictation might be the most selective component of the admission exam, so that differences between groups assessed before and after selection are particularly prominent for this subscale of the BGS. Clearly, longitudinal data are needed to disentangle the relative contributions of selection vs. differential training after

selection to the observed group differences between the two groups.

One clear limitation of our research is the relatively small number of participants, rendering robustness checks of the factor structure difficult. Clearly, follow-up studies with independent samples are needed to confirm the established factor structure. Another limitation of the current study is the lack of other related tests of musical performance or knowledge to validate the test against some standard measures. Note, however, that the test items resemble the kind of tasks students are required to pass during the entrance exam of a top-level music conservatory. Hence, we think that the BGS has high content and face validity. Still, the present study needs to be followed up by more extensive studies that examine the test's psychometric properties in relation to already available measures. We are also aware that item-wise response times are likely also informative about music expertise but those were not recorded individually. Future work may improve upon our model by addressing this issue using a joint response and response latency framework with simultaneous estimation of ability and speed parameters (Prindle, Mitchel, & Petscher, 2016). A further limitation of the current study is the presence of floor and ceiling effects, which might affect statistics and model parameters.

The BGS is designed to assess music expertise. Recent evidence indicates that music expertise cannot be reduced to years of training and formal education attainment. In a study with 10,500 Swedish twins, Mosing and colleagues (2014) found that the amount of music practice was highly heritable, and that associations between musical practice and musical aptitude were substantially correlated with genetic differences. These findings suggest that talent and experience covary, such that music expertise might actually reflect, to a large degree, individual differences in talent. Another study with over 800 pairs of twins by Hambrick and Tucker-Drob (2015) found that genetic effects on music accomplishments were more pronounced among individuals engaging in musical practice, suggesting that talented people are more likely to profit from the experience provided by training. In light of these data, individual differences on the BGS are likely to reflect the outcome of gene-environment interplay.

The BGS complements existing measures by an explicit and exclusive focus on high levels of expertise in Western art music. As such, it complements other measures such as the PROMs and Gold-MSI, which cover a wider range of expertise and often require less time to administer (Zentner & Gingras, 2019). The BGS was developed based on concepts from music theory, and it uses a variety of testing methods derived from musical ear-training practice. Additionally, the BGS requires participants to understand and use musical terminology as well as music notation. To address some of the issues of earlier music aptitude tests, the BGS uses music excerpts from the Western art music tradition from different time periods and with various instrumentations and timbres. The BGS was constructed to provide an objective assessment of music expertise in trained musicians, who represent a comparably small segment of the general population. Given the degree of specialization and depth, its administration also requires more time than most other available measures. Clearly, the BGS is not designed to use as a replacement, but rather as a supplement to other currently available instruments. Researchers are encouraged to use this test to obtain data about participants' music expertise, and to allow for more consistent comparisons of samples across different studies at high levels of music expertise. We hope this tool will benefit future music research in which professional musicians are recruited as research participants.

Author Note

Andreas M. Brandmaier and Elisabeth Wenger share last authorship.

The data from this study are available at https://osf.io/73zym/ and https://osf.io/w59kn. The study was not preregistered.

This study was funded by the Max Planck Institute for Human Development and an Innovation grant by the President of the Max Planck Society given to Ulman Lindenberger.

We are indebted to Steven M. Boker for comments and discussions on the structural equation models. We are grateful to Hanka Theisinger-Hartnack for serving as an independent rater of participants' performance. The authors thank Sarah Polk for her assistance in translating the BGS as well as all participants and student assistants.

References

BLACKING, J., & NETTL, B. (1995). *Music, culture, and experience: Selected papers of John Blacking.* Chicago, IL: University of Chicago Press.

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

BROD, G., & OPITZ, B. (2012). Does it really matter? Separating the effects of musical training on syntax acquisition. *Frontiers in Psychology, 3*, 543. https://doi.org/10.3389/fpsyg.2012.00543

CARROLL, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Oxford, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

CARSON, A. D. (1998). Why has musical aptitude assessment fallen flat? And what can we do about it? Journal of Career Assessment, 6(3), 311–327. https://doi.org/10.1177/106907279800600303

CHAMORRO-PREMUZIC, T., & FURNHAM, A. (2007). Personality and music: Can traits explain how people use music in everyday life?. *British Journal of Psychology, 98*(2), 175–185. https://doi.org/10.1348/000712606X111177

CHIN, T. C., 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*(3), 376–399. https://doi.org/10.1525/mp.2018.35.3.376

CHIN, T., & RICKARD, N. S. (2012). The music USE (MUSE) questionnaire: An instrument to measure engagement in music. *Music Perception, 29*(4), 429–446. https://doi.org/10.1525/mp.2012.29.4.429

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*(2), 117–126. https://doi.org/10.1037/pmu0000213

GORDON, E. (1967). The musical aptitude profile. *Music Educators Journal, 53*(6), 52–54. https://doi.org/10.2307/3390915

GOTTFRIED, T., THOMPSON, G., ELEFANT, C., & GOLD, C. (2018). Reliability of the music in everyday life (MEL) scale: A parent-report assessment for children on the autism spectrum. *Journal of Music Therapy, 55*(2), 133–155. https://doi.org/10.1093/jmt/thy002

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

HAMBRICK, D. Z., & TUCKER-DROB, E. M. (2015). The genetics of music accomplishment: Evidence for gene–environment correlation and interaction. *Psychonomic Bulletin and Review, 22*(1), 112–120.

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

HOOPER, D., COUGHLAN, J., & MULLEN, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60.

KANG, R., NIMMONS, G. L., DRENNAN, W., LONGNION, J., RUFFIN, C., NIE, K., ET AL. (2009). Development and validation of the University of Washington Clinical Assessment of Music Perception test. *Ear and Hearing, 30*(4), 411–418. https://doi.org/10.1097/AUD.0b013e3181a61bc0

KIRCHBERGER, M. J., & RUSSO, F. A. (2015). Development of the adaptive music perception test. *Ear and Hearing, 36*(2), 217–228. https://doi.org/10.1097/AUD.0000000000000112

KREUTZ, G., SCHUBERT, E., & MITCHELL, L. A. (2008). Cognitive styles of music listening. *Music Perception, 26*(1), 57–73. https://doi.org/10.1525/mp.2008.26.1.57

LAW, L. N., & 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

LIM, C. R., HARRIS, K., DAWSON, J., BEARD, D. J., FITZPATRICK, R., & PRICE, A. J. (2015). Floor and ceiling effects in the OHS: An analysis of the NHS PROMs data set. *BMJ Open, 5*(7), e007765. https://doi.org/10.1136/bmjopen-2015-007765

LONSDALE, A. J., & NORTH, A. C. (2011). Why do we listen to music? A uses and gratifications analysis. *British Journal of Psychology, 102*(1), 108–134. https://doi.org/10.1348/000712610X506831

MAS-HERRERO, E., MARCO-PALLARES, J., LORENZO-SEVA, U., ZATORRE, R. J., & RODRIGUEZ-FORNELLS, A. (2013). Individual differences in music reward experiences. *Music Perception, 31*(2), 118–138. https://doi.org/10.1525/mp.2013.31.2.118

MOSING, M. A., MADISON, G., PEDERSEN, N. L., KUJA-HALKOLA, R., & ULLÉN, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science, 25*, 1795–1803. https://doi.org/10.1177/0956797614541990

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

Originally published in: Music Perception, 38(4), 2021, p. 420

NAGEL, F., KOPIEZ, R., GREWE, O., & ALTENMÜLLER. E. (2007) EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods, 39*, 283–290. https://doi.org/10.3758/BF03193159

NIARCHOU, M., SATHIRAPONGSASUTI, J. F., JACOBY, N., BELL, E., MCARTHUR, E., STRAUB, P., ET AL. (2019). Unravelling the genetic architecture of musical rhythm. *bioRxiv*. DOI: 10.1101/836197

OLLEN, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings* (Unpublished doctoral dissertation). Ohio State University.

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*(1),58–75. https://doi.org/10.1196/annals.1284.006

PRINDLE, J. J., MITCHELL, A. M., & PETSCHER, Y. (2016). Using response time and accuracy data to inform the measurement of fluency. In K. Cummings & Y. Petscher (Eds.), *The fluency construct* (pp. 165–186). New York: Springer.

RENTFROW, P. J., & GOSLING, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology, 84*(6), 1236. https://doi.org/10.1037/0022-3514.84.6.1236

RENTFROW, P. J., GOLDBERG, L. R., & LEVITIN, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology, 100*(6), 1139–1157. https://doi.org/10.1037/a0022406

ROSSEEL, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

SAARIKALLIO, S. H. (2008). Music in mood regulation: Initial scale development. *Musicae Scientiae, 12*(2), 291–309. https://doi.org/10.1177/102986490801200206

SAARIKALLIO, S., GOLD, C., & MCFERRAN, K. (2015). Development and validation of the Healthy-Unhealthy Music Scale. *Child and Adolescent Mental Health, 20*(4), 210–217. https://doi.org/10.1111/camh.12109

SCHLAUG, G. (2001). The brain of musicians: A model for functional and structural adaptation. *Annals of the New York Academy of Sciences, 930*(1), 281–299. https://doi.org/10.1111/j.1749-6632.2001.tb05739.x

SEASHORE, C. E. (1919). *Manual of instructions and interpretations for measures of musical talent*. United Kingdom: Columbia Graphophone Company.

SEASHORE, C. E., LEWIS, D., & SAETVEIT, J. G. (1956). *Seashore Measures of Muiscal Talents*. Oxford, England: Psychological Corporation.

SPITZER, J. B., MANCUSO, D., & CHENG, M. Y. (2008). Development of a clinical test of musical perception: Appreciation of music in cochlear implantees (AMICI). *Journal of the American Academy of Audiology, 19*(1), 56-81. https://doi.org/10.3766/jaaa.19.1.6

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

ULLÉN, F., MOSING, M. A., HOLM, L., ERIKSSON, H., & MADISON, G. (2014). Psychometric properties and heritability of a new online test for musicality, the Swedish Musical Discrimination Test. *Personality and Individual Differences, 63*, 87–93. http://dx.doi.org/10.1016/j.paid.2014.01.057

UYS, M., & VAN DIJK C. (2011). Development of a music perception test for adult hearing-aid users. *South African Journal of Communication Disorders, 58*, 19–47. https://doi.org/10.4102/sajcd.v58i1.38.

VANSTONE, A. D., WOLF, M., POON, T., & CUDDY, L. L. (2016). Measuring engagement with music: Development of an informant-report questionnaire. *Aging and Mental Health, 20*(5), 474–484. https://doi.org/10.1080/13607863.2015.1021750

VON OERTZEN, T., BRANDMAIER, A. M., & TSANG, S. (2015). Structural equation modeling with Ωnyx. Structural Equation Modeling: *A Multidisciplinary Journal, 22*(1), 148–161. https://doi.org/10.1080/10705511.2014.935842

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

WERNER, P. D., SWOPE, A. J., & HEIDE, F. J. (2006). The music experience questionnaire: Development and correlates. *The Journal of Psychology, 140*(4), 329–345. https://doi.org/10.3200/JRLP.140.4.329-345

WING, H. D. (1948). *Manual for Standardized Test of Musical Intelligence*. Sheffield, England: City of Sheffield Training College.

WILLIAMS, B., ONSMAN, A., & BROWN, T. (2010). Exploratory factor analysis: A five-step guide for novices. Australasian *Journal of Paramedicine, 8*(3), 1–13. http://doi.org/10.33151/ajp.8.3.93

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. https://doi.org/10.1177/0022429418754845

WONG, P. C., SKOE, E., RUSSO, N. M., DEES, T., & KRAUS, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*(4), 420–422. https://doi.org/10.1038/nn1872

ZENTNER, M., & GINGRAS, B. (2019). The assessment of musical ability and its determinants. In J. Rentfrow & D. Levitin (Eds.), *Foundations in music psychology: Theory and research.* Cambridge, MA: MIT Press.

Originally published in: Music Perception, 38(4), 2021, p. 421