ABSTRACT

Title of Dissertation:	TOWARDS A COMPREHENSIVE MODEL OF MUSICAL ABILITY
	Brooke M. Okada, Doctor of Philosophy, 2018
Dissertation directed by:	Dr. L. Robert Slevc Department of Psychology

Over the past century, multiple tests measuring musical ability have been developed, and research has been investigating individual differences in musical ability to answer questions about the components of musical ability and their dissociations in amusic patients, the innate vs. acquired nature of musical skill, and the potential transfer from musical training to other abilities. However, there has been little consensus on what exactly constitutes musical ability and how to best measure this construct. Previous research has used a variety of tasks assessing mainly perceptual skills (e.g., same/different judgments in sequentially presented melodies), and outcomes from these tasks range from single indices (e.g., pitch ability) to composite scores from multiple tasks (e.g., pitch, rhythm, loudness, timbre). The current study uses individual differences data from 15 representative musical ability tasks (including perception and production measures) to assess the unity and diversity of musical ability, and uses the resulting comprehensive latent measure of musical ability to evaluate previously theorized links between musical ability and individual differences in musical experience, working memory, intelligence, personality factors, and socio-economic status. Results from latent variable model comparisons suggest that musical ability is best represented by related but separate pitch, timing, perception, and production factors. Consistent with previous research, a latent measure of musical ability was positively related to musical training, working memory, and intelligence; in contrast, musical ability was not related to openness to experience or socio-economic status.

TOWARDS A COMPREHENSIVE MODEL OF MUSICAL ABILITY

by

Brooke M. Okada

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor L. Robert Slevc, Chair
Professor Michael Dougherty
Professor Kenneth Elpus
Dr. Jared Linck
Dr. Tracy Tomlinson

Acknowledgements

First and foremost, I would like to express my appreciation for my advisor Bob Slevc for his guidance, patience, and support over the past few years. His advice regarding research and teaching have been invaluable. I would also like to thank my committee members for their helpful comments and suggestions on my work, my lab mates for lively discussion and snacks, and all of the undergraduate research assistants who helped with data collection. My sincere thanks also goes to my friends here on the East Coast and back in California for the late-night talks, laughs, and the strength to keep on keeping on. Last but not least, I would like to thank my family, especially my parents, Ed and Elaine Okada, who encouraged me to move across the country to pursue my education and for supporting me every step of the way.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Music has been around for centuries: the oldest instrument that has been uncovered is a bone flute thought to be more than 30,000 years old (Conard, Malina, & Munzel, 2009), and humans developed the capacity to vocalize and possibly sing thousands of years prior (Dediu & Levinson, 2013). Psychology has a long history of working to measure individual differences in cognitive abilities (as well as some controversy over what those individual differences mean, e.g., Murdoch, 2007). Although less publicized than intelligence testing, measures of musical ability have a similar history. Furthermore, individual differences in music (as well as other cognitive functions) are again attracting growing interest, fueled both by interest in cognitive transfer and understanding the nature of perceptual and cognitive abilities.

Musical ability has been conceptualized in a number of ways over the past few decades. In fact, in a study where participants completed the phrase "Musical ability is:" multiple categories emerged, such as generative skills (e.g., playing an instrument, singing, and composing music), receptive responses (e.g., listening, understanding, appreciating, and evaluating music music), as well as the origins of music ability (e.g., its innate and learned aspects) (Hallam & Prince, 2003). Although most researchers acknowledge these distinctions, previous terminology has mainly emphasized the innate nature of musical ability, which has been given different terms by different researchers. For example, musical capacity has been defined as the "inborn psycho-physic and mental capacities distinguished from skills acquired in training" (Seashore, 1915, p. 129), musical ability has been defined as the "potential for learning music before formal training and achievement" (Law & Zentner, 2012, p. 2), and musical aptitude is thought to be the "natural music abilities or innate potential to succeed as a musician" (Schellenberg & Weiss, 2012, p. 499).

The goal of the earliest test of musical ability was to assess children's innate aptitude for music lessons and whether or not they should pursue a career in the arts (i.e., Seashore's Measures of Musical Talents, Seashore, 1919; 1938; 1960). It is worth noting that Seashore's tests were actually developed to investigate the heredity of musical ability, which dovetailed with his interest in eugenics (e.g., see Koza, 2007; Shuter, 1966; Stanton, 1922). Currently, the most prevalent tests used are the series of musical ability tests created by Edwin Gordon (1965; 1989). These tests are usually used in the field of music education to determine music students' strengths and weaknesses, determine who may have the aptitude for long-term music lessons, and are also typically used as a covariate in research examining effects of music training on non-musical abilities (i.e., to control for the issue that some participants may be predisposed to be good at music, Bugos, Perlstein, McCrae, Brophy, & Bedenbaugh, 2007; Bugos & Kochar, 2017). This approach makes sense if musical ability is an innate skill; indeed, Carl Seashore argued that "musical talent … is a gift of nature -- inherited, not acquired" (Seashore, 1915, p. 129), and there is evidence that music ability is heritable (Mosing et al., 2014; Tan, McPherson, Peretz, Berkovic, & Wilson, 2014).

However, others suggest that musical ability is primarily an experience-based effect. For example, Shinichi Suzuki claimed that "there is no such thing as an innate aptitude for music" (Hermann, 1981, p. 137), and "any child has seeds of ability which can be nurtured as far as the capacity of the brain will allow" (Suzuki, 1981, p. 2). Previous studies suggest that experience does play an important role in acquiring musical expertise (e.g., Chen et al., 2008; Ericsson, Krampe, & Tesch-Romer, 1993), and that the amount of musical experience predicts musical ability (e.g., Peretz et al., 2013; Slevc, Jaeggi, Buschkuehl, & Davey, 2016) although the directional relationship is unclear. Additionally, current work has tended to focus more on experience, in part because of relationships between musical training and other cognitive abilities like language (Slevc & Miyake, 2006), executive functions (Okada & Slevc, 2018; Slevc et al., 2016), and intelligence (Schellenberg, 2004).

More recent work acknowledges that one's musical skill may be determined both by an innate potential to succeed in music and experience/practice in musical training, and attempts to measure individual differences in musical skills reflecting both innate and learned aspects. Again, terminology for this varies from study to study. For example, musical competence has been defined as "listeners' ability to perceive, remember, and discriminate musical melodies and rhythms" and is "meant to be neutral with respect to the relative roles of nature and nurture" (Swaminathan & Schellenberg, 2018, p. 1), and musical ability is conceptualized as performance on a set of perceptual judgments (Law & Zentner, 2012; Wallentin et al., 2010). Like most previous studies, the current study does not attempt to disentangle innate versus learned musical abilities, but assumes that musical ability exists (i.e., reflecting some combination of pre-dispositional and experiential factors) and that tests measuring it are capturing variation specific to the various components of musical ability.[1] In sum, although different aspects constituting musical ability have been studied extensively and multiple tests measuring musical ability have been developed, conclusions about musical ability are limited by inconsistent definitions and the use of only single (or a couple) perceptual measures of musical ability.

Aspects of Musical Ability

Models of music processing vary in how they conceptualize various musical factors. Outlined below are relevant musical terms that will be used to discuss the models. (Note that following most studies on music, this study focuses solely on Western tonal music, although it will clearly be important to pursue these issues in non-Western musical traditions as well).

The basic unit of music is pitch. The perception of pitch is based on the periodicity of soundwaves, with sounds with greater frequency (measured in Hertz (Hz), or cycles per second) sounding higher pitched than those with lower Hz. Multiple sequential pitches form melodies, and multiple pitches played simultaneously (i.e., chords) form harmony (see Oxenham, 2013, for a review). Another relevant aspect of music is timing, which is represented in tempo, meter, and rhythm. In music, tempo refers to the rate or pace of the music, and is also conceptualized as the time interval between beats. Meter refers to the organization, or

[1] Here, musical ability is assumed to be variable and influenced by experience even though it is described as a fixed ability (due to measurements only capturing a fixed snapshot of one’s ability).

regular pattern of repeated beats (usually indicated by a time signature such as 4/4). Lastly, rhythm deals with the temporal organization of different patterns of notes (noises) and rests (silences) (see McAuley, 2010, for a review).

Most models of music processing distinguish between at least some of these aspects of music. Most commonly, models tend to distinguish between pitch and timing/rhythmic processing and/or between perception and production of music, detailed below.

Pitch vs. Timing Processing

Most models of musical processing posit that musical processing is modular, and separate from other general types of auditory processing (e.g., language processing, Peretz & Coltheart, 2003, but see Koelsch et al., 2002; Koelsch & Siebel, 2005; Koelsch, 2011). Within Peretz and Coltheart's (2003) modular musical processing model (see Figure 1), there are thought to be separate pitch and timing processing modules, which operate in parallel with one another (Peretz, 2001; Peretz & Coltheart, 2003). The pitch module processes pitch, melodic contour, and intervals within a melody or harmony, and the temporal module processes the timing component of music, which encompasses meter and rhythm processing (Peretz & Coltheart, 2003).

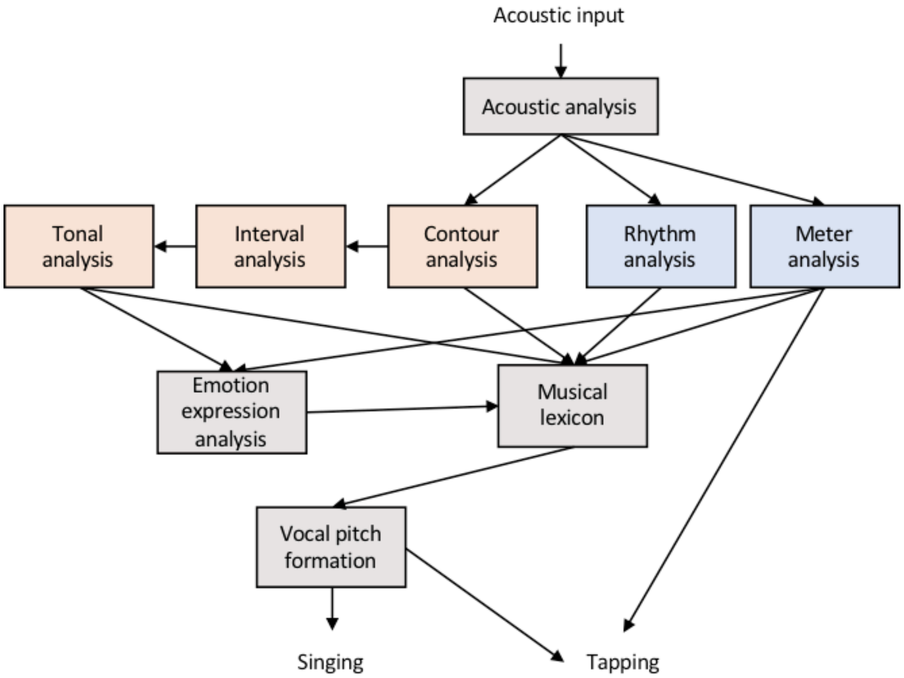


Figure 1. Model of music processing, adapted from Peretz & Coltheart (2003). Peach-colored boxes specify modules specific to pitch processes and blue boxes specify modules specific to timing processes.

Evidence for the separability between pitch and timing abilities comes from multiple studies on congenital amusia, a lifelong deficit in music processing that is estimated to affect around 2-4% of the population (Peretz, Champod, & Hyde, 2003; Peretz, Cummings, & Dubé, 2007; Peretz et al., 2013), as well as evidence from acquired amusia (i.e., amusia resulting from stroke or other brain injury). Even though individuals with congenital amusia have normal hearing abilities, they have trouble recognizing familiar melodies, recognizing poor singing, and do not show sensitivity to dissonant or out of key chords (Peretz et al., 2007). This “tone deafness” characterizes most amusics, whose processing deficits show a dissociation between impaired musical pitch processing, but intact rhythm processing (Ayotte, Peretz, & Hyde, 2002; Peretz et al., 2007). The opposite pattern has also been found, with some amusics having intact pitch processing, but deficient rhythm processing (e.g., “beat deafness,” or poor discrimination between differing rhythms and poor reproduction of rhythms, Alcock, Passingham, Watkins, & Vargha-Khadem, 2000; Phillips-Silver et al., 2011). The neuropsychological literature, as a whole, suggests fractionated musical abilities; however, it is hard to draw firm conclusions on patient data alone because it is unknown if the structure of these abilities are mirrored in the general population. Furthermore, it is unknown if amusic patients have developed strategies to overcome some of their deficits or if the patients included in these studies are at the extreme ends (i.e., only those with very noticeable deficits self-select into these studies). In fact, recent evidence suggests that with pitch training, amusic patients can show and sustain improvements in pitch and melody discrimination ability, leading them to no longer be classified as amusic (Whiteford & Oxenham, 2018). And while tests of musical ability have been developed to find cases of amusia (e.g., The Montreal Battery of Amusia (MBEA), Peretz, 2001; Peretz et al., 2003, and the Montreal Battery of Evaluation of Musical Abilities (MBEMA), Peretz et al., 2013), most are intended for normally functioning participants.

Perception vs. Production Processing

Even when looking at the same construct (e.g., pitch or timing ability), the demands are different with perception and production. In perception, the goal is to extract information regarding pitch and timing from a continuous auditory signal, and in production, the goal is to translate an intended pitch and/or rhythm into motor movements to produce an auditory signal.

Most tasks measuring musical ability focus primarily on perceptual abilities. To assess musical perception, participants are usually tasked in perceiving whether two stimuli are the same or different, or they complete some sort of auditory discrimination threshold task (e.g., Gordon, 1965; 1989; Peretz et al., 2003; 2012; Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010; Law & Zentner, 2012). During musical production, tasks vary from having the participant sing and imitate different pitches (e.g., Alcock et al., 2000; Pfordresher & Brown, 2007; Pfordresher, Brown, Meier, Belyk, & Liotti, 2010) to tapping along with metronome beeps or songs (e.g., Dalla Bella et al., 2017; Fujii & Schlaug, 2013; Iverson & Patel, 2008). Interestingly, some amusic patients have shown both intact ability for pitch perception and pitch production, but showed deficits in both rhythm perception and rhythm production (Alcock et al., 2000). But in other cases, there seems to be a dissociation between perception and production abilities for both pitch and timing. For example, Dalla Bella and colleagues (2017) administered four perceptual tasks (duration discrimination, anisochrony detection with pure tones, anisochrony detection with music, and the Beat Alignment Test) and five production timing tasks (unpaced tapping, paced tapping to isochronous rhythm, paced tapping to music, synchronization-continuation, and adaptive tapping where the tempo fluctuated), and found that although some of the perception and production tasks were significantly correlated with one another, many were not. A similar mixed pattern of results was found in another study assessing timing perception and production, suggesting a dissociation between timing perception and production abilities (Fujii & Schlaug, 2013). These results fall in line with similar dissociations in pitch perception and pitch production abilities (e.g., Dalla Bella, Berkowska, & Sowinski, 2011; Loui, Guenther, Mathys, & Schlaug, 2008), but other research has shown that pitch perception and pitch production performance scale with one another (Amir, Amir, & Kishon-Rabin, 2003). Further, Norris (2000) found that performance on a pitch production test correlated with only some pitch perception tasks and varied based on the age group studied. Although this literature suggests that perception and production abilities may be separable, it is difficult to draw firm conclusions given the mixed findings and that most previous tests of musical ability only measure musical perception abilities and do not consider musical production abilities.

The Role of Perceptual Acuity

There is also the possibility that any individual differences in musical ability may be due to more general perceptual abilities/auditory acuity. Western tonal harmony is confined to discrete pitches that differ by 100 cents (or a semitone); however, there is variation in performance on more fine grained perceptual acuity tasks (e.g., pitch discrimination tasks that find thresholds smaller than 100 cents and duration discrimination tasks; Kidd, Watson, & Gygi, 2007). Previous work suggests that some general auditory acuity tasks may not be correlated with more complex musical timing perception tasks (e.g., Dalla Bella et al., 2017); however, other work finds that musical experience/ability is associated with advantages in early perceptual processes (e.g., the brainstem frequency following response; Bidelman, Gandour, & Krishnan, 2010; Kraus & Chandrasekaran, 2010).

Abridged History of Musical Ability Tests

Some previous tests of musical ability correspond to these theoretical divisions (e.g., pitch/timing distinction), but most contain subtests that the authors themselves deemed important, which do not directly appeal to theory (see Table 1 for a summary). For example, Seashore's Measures of Musical Talents (Seashore, 1919) contains several perceptual subtests tapping different aspects of musical ability and were later revised to contain subtests of pitch, rhythm, loudness, time, timbre, and tonal memory (Seashore, 1960). Seashore argued that each subtest measured one aspect of musical ability individually, and that these measures should not be combined into one individual musical aptitude score, but should be considered as a whole to describe one's overall musical aptitude profile.

Wing's Tests of Musical Aptitude (Wing, 1948; later revised in 1962) rejected Seashore's view that each subtest should not be combined to form one musical aptitude score, and developed a test consisting of seven perceptual subtests, whose scores were combined to yield an overall score of musical ability. The first three tests were tests of pitch perception ability, and the last four were subjective tests of judgments of musicality (i.e., choosing which of two performances of the same piece had the better: rhythmic accent, harmonizations, loudness, or phrasing).

Edwin Gordon first developed the Musical Aptitude Profile (MAP, Gordon, 1965), which focused on audiation, which he defined as "the foundation of music aptitude…the ability to hear and to comprehend music for which the sound is not physically present (as in recall), is no longer physically present (as in listening), or may never had been physically present (as in creativity and improvisation)" (Gordon, 1995, p. 8). The MAP contains seven subtests tapping tonal imagery (same/different melody and harmony tasks), rhythm imagery (same/different tempo and meter tasks), and musical sensitivity (determining which passage had the best phrasing, balance, and style). He later devised the Primary Measures of Music Audiation (PMMA), Intermediate Measures of Music Audiation (IMMA), and Advanced Measures of Music Audiation (AMMA), which all contained rhythmic and melodic same/difference judgments of various lengths suitable for different age groups (Gordon, 1989; 2001). Like the other previous tests, Gordon's assessed only musical perceptual abilities.

The Montreal Battery of Evaluation of Amusia (MBEA, Peretz, 2001; Peretz et al., 2003) was designed to evaluate perceptual music deficits in adults, (rather than to measure individual differences in a typical population) and contains six subtests measuring pitch and rhythm abilities, in which same/different judgments are made. In order to evaluate music deficits in children, the Montreal Battery of Evaluation of Musical Abilities (MBEMA, Peretz et al., 2013) was developed, with shorter stimuli and an overall shorter testing time. The MBEMA also contains four similar subtests measuring pitch and rhythm abilities, and the Abbreviated MBEMA contains two. Both the MBEA and MBEMA were designed to measure one's overall/total musical ability score and to determine whether someone is amusic (i.e., receiving a score below 2SD of the mean).

A similar test designed to measure variation in musical ability is the Musical Ear Test (Wallentin et al., 2010), which consists of rhythmic and melodic same/different judgments. Finally, a more recently developed test of musical perception ability is the Profile of Music Perception Skills (PROMS, Law & Zentner, 2012), which contains subtests of same/different judgments of multiple facets of music including melody, rhythm, rhythm-to-melody (i.e., recognizing a rhythmic pattern when it is part of a melody), accent, tempo, pitch, timbre, tuning, and loudness. Shorter versions of the PROMS have also been created and validated to facilitate testing time constraints. The Shortened PROMS (PROMS-S) includes the same subtests as the PROMS, but with fewer items (Zentner & Strauss, 2017), the Brief PROMS contains only the melody, accent, tempo, and tuning subtests (see Kunert, Willems, & Hagoort, 2016 for a psychometric evaluation of the Brief PROMS), and the Mini PROMS is an abbreviated version of the Brief PROMS (Zentner & Strauss, 2017). Table 1 summarizes the various aspects of musical abilities assessed by these different tests. All assume at least some different aspects of musical abilities, and most include at least distinct tests for pitch-based and timing-based abilities. Interestingly, none of these musical ability tests contain tasks measuring production. Furthermore, the PROMS Pitch and Tuning subtests (Law & Zentner 2012; Zentner & Strauss, 2017) are the only ones that look at more fine grained auditory acuity (i.e., at differences smaller than a semitone). Notably, some of these tests (i.e., Gordon's ability tests) require licensing and purchase, whereas others are made available for free (or by request) (e.g., Law & Zentner, 2012; Peretz et al., 2003; Peretz et al., 2013; Wallentin et al., 2010; Zentner & Strauss, 2017).

Table 1

Summary of Selected, Commonly-Used Previous Musical Ability Tests

Subtests	Seashore	Wing	Gordon (MAP)	Gordon (AMMA)	Peretz (MBEA, MBEMA)	Wallentin (MET)	Law (PROMS, PROMS-S)
Melody	x	x		x	x	x	x
Rhythm	x			x	x	x	x
Pitch	x	x					x
Loudness	x	x					x
Accent		x					x
Tempo			x				x
Timbre	x						x
Phrasing		x	x				
Duration/Time	x						
Chord Analysis		x					
Harmony		x	x				
Meter			x				
Style			x				
Balance			x				
Tuning							x

Note. Table adapted from Law, 2012

Musical Ability's Relationship with Other Abilities

In addition to assessing a model of musical ability, another aim of this study was to assess musical ability's relationship with other factors that are likely relevant to musical ability. These include musical training, working memory, intelligence, personality (specifically, openness to experience), and socio-economic status (e.g., Slevc et al., 2016; Swaminathan et al., 2017; Swaminathan & Schellenberg, 2018). By assessing these relationships with a more comprehensive battery of musical ability measures, these comparisons aimed to provide a clearer picture of the robustness of these relationships.

One variable commonly investigated in tandem with musical ability is musicianship, or amount of musical training (i.e., formal music lessons). Musical training's relationship with various non-musical cognitive abilities has received much interest, stemming from the possibility of benefits (i.e., transfer effects) of music lessons (for reviews, see Benz, Sellaro, Hommel, & Colzato, 2015; Okada & Slevc, in press; Schellenberg & Weiss, 2013). Musical ability has been shown to be related to musical training (Swaminathan et al., 2017; Swaminathan & Schellenberg, 2018) although the directionality of the relationship is uncertain. Nonetheless, these results suggest that those with high musical ability skills may pursue and persist longer in musical training and/or musical training could enhance musical ability. Supporting the latter, a recent longitudinal study demonstrated that children randomly assigned to musical training showed higher performance on a tonal discrimination task than children in sports training and children in a no contact control group (Habibi, Damasio, Ilari, Sachs, & Damasio, 2018).

Whether examining relationships with musical training or performance on musical ability measures, a concern is that any observed relationships might reflect shared reliance on domain-general cognitive abilities rather than (or in addition to) musical ability per se. For example, during a same/different judgment trial, one needs to hold in mind the first sequence in order to compare it to the second sequence. And in order to process and make sense of pitch and timing, the notes must be compared to a reference frame (e.g., harmony or meter) (Wallentin et al., 2010). These processes may necessitate the use of working memory as well as processes like pattern recognition, akin to fluid intelligence. In fact, studies have shown that musical ability is related to performance on short-term memory measures such as the digit span (Swaminathan & Schellenberg, 2018; Wallentin et al., 2010), working memory updating abilities (Okada & Slevc, 2018; Slevc et al., 2016) and intelligence (Swaminathan & Schellenberg, & Khalil, 2017).

In recent research, individual differences in personality traits (specifically, openness to experience) have been used to predict various aspects related to music ability. Openness to experience is thought to be indicative of intellect and sensitivity to artistic, aesthetic experiences as well as curiosity and imagination (John et al., 2004). Indeed, openness to experience has been shown to predict who takes music lessons (Corrigall, Schellenberg, & Misura, 2013), amount of music practice (Butkovic, Ullen, & Mosing, 2015), musical competence (Swaminathan & Schellenberg, 2018), and musical sophistication (Greenberg, Müllensiefen, Lamb, & Rentfrow, 2015). Openness to experience has also been found to predict performance on a musical ability test; however, this relationship was mediated by musical training (Thomas, Silvia, Nusbaum, Beaty, & Hodges, 2016). These findings suggest that those higher in openness may be more likely to try a musical instrument and/or seek out more musical experiences, either of which could explain the relationships found between openness and musical ability (Thomas et al., 2016; Swaminathan & Schellenberg, 2018).

Of course, musical training requires a certain amount of time and resources; correspondingly, measures of socioeconomic status (SES) have also been shown to correlate with musical participation (Corrigall et al., 2013; Elpus, 2013; Kaushal, Magnuson, & Waldfogel, 2011; Norton et al., 2005; Southgate & Roscigno, 2009, but see Okada & Slevc, 2018; Slevc et al., 2016) and musical ability (Swaminathan et al. 2017; Swaminathan & Schellenberg, 2018). Since those with higher SES have more opportunity for musical training and other musical activities, this study sought to investigate the relationships between SES and musical ability.

While these relationships would be interesting themselves, the ways in which specific musical ability measures rely on different perceptual and cognitive abilities clearly hampers the ability to observe musical/cognitive relationships more generally.

Current Study Objectives

The goal of this study was to investigate various aspects of musical ability based on theoretically proposed distinctions and distinctions found in previous individual differences work, and to develop a theory/model that specifies a more nuanced picture of what constitutes musical ability. Previous studies have advocated for a range of different factors (including pitch, harmony, timing, expressiveness, etc.), and previously developed tests of musical ability have emphasized factors that the authors deemed most important (e.g., Law & Zentner's (2012) test of musical ability taps nine characteristics of sound, and Wing's (1919; 1960) tests make the distinction between tests of perceptual ability and tests assessing judgment of the best expression/musicality within excerpts). Based on the previous literature regarding musical ability, the models assessed here focused on the most commonly used distinctions already present in the field: pitch and timing. Furthermore, most current tests only measure perception abilities, where the trials consist of listening to two musical stimuli and judging whether they are the same or different. However, because perception and production abilities have been shown to differ within certain populations, tests of production ability were also included. This study used a latent variable approach to conduct a large individual differences study with multiple tasks measuring pitch perception, pitch production, timing perception, and timing production. Model comparisons were used to ascertain whether individual differences in these abilities fit the models hypothesized by the literature. This study also examined the extent of the relationship between pitch and timing abilities, perception and production abilities, and the relationship between finer-grained auditory acuity and processing of more complex music.

A secondary goal of this study was to assess if and how musical ability relates to other non-musical abilities and factors. This provided a more nuanced understanding of how different aspects of musical ability relate to non-musical abilities such as working memory and fluid intelligence, which may help clear up current controversial findings about the relationship between musical and non-musical abilities (e.g., Moreno et al., 2011; Okada & Slevc, 2018). In addition, this provided a better understanding of if and how performance on these musical ability assessments reflects other situational and cognitive factors that are likely (albeit somewhat controversially) related to musical abilities (e.g., musical training, personality, and SES).

Chapter 2: Method

Pre-registration

Due to the importance of research reproducibility and openness (e.g., Nosek, Spies, & Motyl, 2012), this project was pre-registered before data was collected on the Open Science Framework (http://opensciencefranework.org/), and a frozen, non-editable registration form is available at https://osf.io/jwyhu (see, e.g., van't Veer & Giner-Sorolla, 2016 for the importance of pre-registration). All of the methods, exclusion criteria, and analyses were conducted as specified in the pre-registration form except where noted.

Participants

167 participants (total *N* = 165 after two exclusions[2]) were recruited from the University of Maryland's undergraduate research pool and received class credit for participation. A target sample size of 150 participants (or as close as possible by June, 2018[3]) was set a priori based on similar participant numbers in other studies

[2] One participant asked for their data to be removed from the study after participation, and one other participant's data was removed because they scored above the pre-registered cut off score of 32 on the *Revised American Academy of Otolaryngology-Head & Neck Surgery Five-Minute Hearing Test*, signifying a difficulty in communication and requiring a hearing test/solution (Kochkin & Bentler, 2010; Koike, Hurst, & Wetmore, 1994).

[3] Data was collected until the end of the Spring semester, and so the final sample size was slightly larger than the targeted *N* of 150.

examining individual differences in cognitive processes (120 in Conway, Cowan, Bunting, Therriault, & Minkoff, (2002); 133 in Engle, Tuholski, Laughlin & Conway (1999); 137 in Miyake et al. (2000); 215 in Shipstead, Lindsey, Marshall, & Engle, 2014; 127 in von Bastian & Druey, 2017; 92 in von Bastian & Oberauer, 2013). Participants were on average 19.86 years of age (*SD* = 1.69), with a range of 18-29 years. Ninety-nine participants (60%) identified as female, 63 (38%) identified as male, 2 identified as non-binary, and 1 identified as trans-male. Although all participants spoke English fluently, some reported other languages as their first, native language. 149 participants reported English as their first language, 5 reported Chinese/Mandarin, 3 reported Korean, and 1 reported each of the following: Gujarati, Igbo, Nepali, Portuguese, Spanish, Tagalog, Tamil, and Vietnamese.

Measures

Additional information on all of the measures used here is available on the Open Science Framework at https://osf.io/mp3u7/.

Musical Measures

Participants completed a battery of 15 musical tasks measuring pitch ability and timing ability (see Table 2 for a summary). Versions of these tasks have all been used previously to measure various aspects of musical perception ability, auditory acuity, or musical production ability. Although some of these tasks come from a single test battery (e.g., the PROMS-S has 3 pitch tasks and 2 timing tasks), each task is grouped and described under its respective heading, i.e., Pitch Perception Measures, Pitch Production Measures, Timing Perception Measures, and Timing Production Measures.

Pitch Perception Measures. These five tasks – PROMS-S Melody subtest, PROMS-S Tuning subtest, PROMS-S Pitch subtest, Chord Analysis, and Pitch Threshold Discrimination – all required perceptual judgments of pitch.

PROMS-S. The shortened version of the Profile of Music Perception Skills (PROMS-S)[4] that has been shown to be reliable and consistent (Zentner & Strauss, 2017) was used to keep testing time to a minimum. All tasks from the PROMS-S battery consisted of same/different judgments (with an equal number of same and different trials, balanced for difficulty) and were administered on LimeSurvey, an online survey application[5] (Law & Zentner, 2012; Zentner & Strauss, 2017). Transcriptions of all musical stimuli from the PROMS-S are included at https://osf.io/mp3u7/. Five subtests from the PROMS-S were administered overall, three of which involve pitch perception and are described individually below. All five subtests followed the same procedure: participants were tasked with determining if two sequences were the same or different. In each trial, participants listened to the

[4] The original PROMS included 9 subtests with 18 items each (9 same trials and 9 different trials, with 3 easy, moderate, and complex trials each) (Law, 2012; Law & Zentner, 2012). To shorten the total testing time, the PROMS-S used only trials meeting the following criteria: difficulty level between 5%-95%, item-to-total correlation of 0.30 or higher on that subtest, within 10% of reliability reported for the original PROMS (Zentner & Strauss, 2017).

[5] All versions of the PROMS are free to use and provided by the authors to administer on LimeSurvey.

first musical stimulus twice, heard the comparison musical stimulus, then indicated whether they thought the comparison musical stimulus was “definitely different,” “probably different,” “I don’t know,” “probably same,” or “definitely same”. Each subtest contained 2 practice trials (one same, one different) to allow participants to familiarize themselves with the task. To measure aspects of pitch perception, the Melody, Tuning, and Pitch subtests from the PROMS-S were administered and are described below.

The PROMS-S Melody subtest consisted of 10 trials with sequences of varying length (9-19 tones) voiced in a harpsichord timbre ranging from C4 (middle C) to A-flat5. Participants were tasked with determining if two melodies were the same or different. Trials ranged in difficulty, with easier trials containing simpler rhythms and different pitches on down beats, and more complex trials containing more notes and accidentals and different pitches on passing notes (see Figure 2 for an example of a different trial). Following Law and Zentner (2012) and Zentner and Strauss (2017), participants’ raw accuracy scores were weighted to include confidence levels by scoring 1 point for a correct “definitely” answer, 0.5 point for a correct “probably” answer, and 0 for all other answers. Then, scores were converted to d-prime (d’), a measure of sensitivity calculated as the difference between the proportion of hits and false alarms (z(Hits) - z(False Alarms))[6]. In the case of a participant achieving floor or ceiling performance (i.e., 0 or 1 for hit rate or false alarm rate), d’ could not be calculated, so these extreme values were replaced. A rate

[6] A d’ score of 0 indicates no sensitivity/no discrimination ability, d’ score of 1 indicates 69% correct for both different and same trials, and higher numbers indicate greater sensitivity.

of 0 was replaced with 0.5/n and a rate of 1 was replaced with (n-0.5)/n, where n was the number of trials (Macmillan & Kaplan, 1985). This effectively increased or decreased performance levels by half a trial and allowed for calculation of d'.

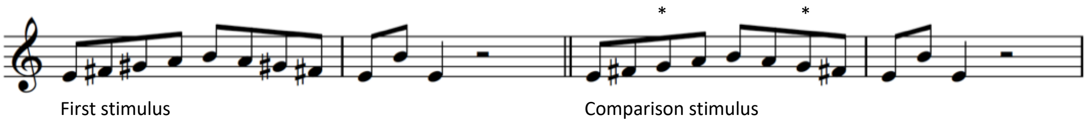


Figure 2. Example of a different stimulus from the Melody subtest (from Law, 2012). The asterisks indicate the changed note within the second sequence.

The PROMS-S Tuning subtest presented participants with a C Major chord (C4, E4, G4, C5) voiced in a piano timbre for 1.5 s. Within the C Major chord, the E was mistuned by a variable amount (0-50 cents) and the participant determined if the comparison chord had the same amount of mistuning as the first chord. The eight trials ranged in difficulty, with easier trials containing larger amounts of mistuning change, and more complex trials containing smaller amounts of mistuning change. Like the Melody subtest, participants' raw accuracy scores were weighted to include confidence levels, then transformed to d'.

The PROMS-S Pitch subtest consisted of pairs of pure tones (2000 ms sinusoids with 100 ms linear on/off ramps) centered around A4, and participants determined if the comparison tone was the same or different (i.e., either lower or higher) pitch than the first tone. The eight trials ranged in difficulty, with easier trials containing larger pitch changes, and more complex trials containing smaller pitch changes. Again, participants' raw accuracy scores were weighted to include confidence levels, and were transformed to d'.

Chord Analysis. In this subtest from Wing's revised Musical Aptitude Test (Wing, 1962), participants listened to 20 musical stimuli voiced by a piano timbre presented in PsychoPy (v 1.82.0, Peirce, 2007; 2009). Each stimulus was either a single note or a chord, ranging from one to six notes (see https://osf.io/mp3u7/ for transcriptions of stimuli). Participants were instructed to determine the number of notes played in each stimulus, and were given four practice trials (containing two, three, one, and four notes) to familiarize themselves with the task before the 20 real trials. Participants' scores were calculated as proportion accuracy, following the procedure used by Wing (1962).

Pitch Threshold Discrimination. In order to measure perceptual acuity in pitch, participants' pitch threshold discrimination was measured using the Pitch Discrimination (Pure Tone) task from the Psychoacoustics Toolbox (Grassi & Soranzo, 2011; Soranzo & Grassi, 2014) designed for Matlab. In this adaptive[7] task, participants completed two blocks of 30 trials each. In each trial, participants heard three serially presented 250-ms pure tones, and determined which was the highest in pitch (3AFC). (This task can also be completed with a 2AFC; however, a 3AFC was chosen because it can more robustly estimate thresholds and using 2AFC are more affected by false alarms (Grassi & Soranzo, 2014). In this 3AFC task, two of the tones were the same and the other was presented at a variable frequency (either below or above the participants discrimination threshold). Based on each participants' performance, the discrimination threshold was calculated using the maximum likelihood estimation procedure, which employs the maximum likelihood estimation and the stimulus selection policy. Each participant's data was fitted to a sigmoid

[7] The adaptive maximum likelihood procedure was chosen over the staircase method because it uses all of the available data and can estimate participants' thresholds more accurately and is faster to administer (Grassi & Soranzo, 2009).

function (logistic used here because it is the most widely used), which represents the proportion of correct responses for varying stimulus/frequency levels. In order to estimate each participant's performance, multiple functions with different midpoints are hypothesized, and the maximum likelihood estimation finds the function that is most like the participant's after each trial. After this, each participants' threshold was calculated by finding the stimulus level that corresponded to the point on the function at which the participant achieved 72.87% correct responses (called the p-target, which was set at 0.7287 following Grassi & Soranzo, 2014). Each participant's discrimination threshold was calculated in both blocks, and the average of the two blocks was used as the final pitch threshold discrimination score.

Pitch Production Measures. Participants completed three pitch imitation tasks – Familiar Song Imitation, Mowrer Test of Tonal Memory, and Melody Imitation – to measure pitch production ability. For all three tasks, participants were recorded using the internal microphone of an Apple iMac through PsychoPy (v 1.82.0, Peirce, 2007; 2009)[8]. Before participants began the Pitch Production tasks, they warmed up with the experimenter by clearing their throat, then completing vocal sweeps (i.e., singing from the lowest note they could hit to the highest, then back down from the highest to the lowest).

[8] Unfortunately, the recordings from the internal microphone of the Apple iMac desktops occasionally recorded heavy amounts of static noise. Because this extra noise made it difficult to extract accurate pitches from the sound files, all trials containing static were discarded. This resulted in multiple utterances discarded due to faulty recording equipment (22.7% of the utterances in the Familiar Song Imitation Task and 12.9% of utterances in the Melody Imitation Task). Furthermore, two participants did not have any singing production data due to experimenter error (when the microphone was not enabled prior to recording). These exclusion criteria were not included in the pre-registration, but were deemed necessary for analysis purposes.

Familiar Song Imitation Task. In this task, following Pfordresher et al. (2010), participants sang four familiar songs (e.g., Happy Birthday; see https://osf.io/mp3u7/ for complete list of song lyrics). First, participants read the song lyrics on the screen, then sang them aloud starting on a note that was comfortable for them. Each of the four songs was presented one at a time. From these four songs, a pre-specified set of 10 intervals[9] in total were analyzed (following those analyzed in Pfordresher et al., 2010). To analyze pitch interval accuracy, each utterance containing the notes of interest was extracted using Audacity (v 1.3.12, Audacity Team, 2010). Then, Praat (v 6.0.19, Boersma & Weenink, 2016) and PraatR (v 2.4, Albin, 2014) were used to find the mean pitch in Hertz (Hz) of each utterance. Intervals (in cents)[10] were calculated with the formula: cents = 1200 x $\log_2(f_2/f_1)$, where f_1 is the frequency (in Hz) of the first note and f_2 is the frequency (in Hz) of the second note. This observed interval was then subtracted from the expected interval (e.g., a perfect fifth is 700 cents) to determine accuracy. Interval jumps greater than 1800 cents (i.e., 1.5 octaves) were assumed to be pitch tracking errors and were discarded. Four participants reported not knowing at least one of the songs, so their accuracy scores reflect ability for the trials they did complete. Participants' accuracy scores were averaged across all trials.

[9] In Pfordresher et al., 2010, six songs total were sung by participants; however, only four of the songs were used here to keep testing time to a minimum. In these four songs, Pfordresher et al. (2010) analyzed 11 total intervals. Here, one interval from Happy Birthday (the minor sixth) was not analyzed because multiple people did not sing the lyric containing part of the minor sixth (i.e., a name). This was perhaps due to the fact that blank lines were shown in place of a name.

[10] Cents were calculated (instead of Hz) because the same interval (i.e., P4) has a smaller Hz difference in higher ranges than lower ranges, and cents uses a log transform to equally scale these differences.

Mowrer Test of Tonal Memory. In this task taken from Mowrer (1994), participants heard melodies of varying length (five to nine notes) played by a piano timbre and were asked to sing and reproduce the melody using the syllable "/da/" for each note. To allow participants to familiarize themselves with the task, a practice trial was given (five notes long), and the experimenter provided feedback if the participant was not singing "/da/" for each individual note. Participants then completed seven trials. Following Mowrer (1994), each trial was scored as either correct or incorrect, so accuracy scores ranged from zero to seven. A trial was counted as incorrect if any of the pitches were incorrect, including if some of the pitches were missing. Scores used in the final analyses were ratings done by the author blind to the participants' other scores; however, to ensure the reliability of these scores, a subset of the recordings (19/165) were also rated by a UMD DMA Candidate in Vocal Performance. Kappa of .95 was achieved, showing excellent inter-rater reliability (Cohen, 1960; Hallgren, 2012).

Melody Imitation Task. Again following Pfordresher et al. (2010), participants imitated 24 five-note sequences[11] voiced by a synthesized voice by singing each pitch on "/da/". First, participants decided whether they felt most comfortable imitating lower pitched notes or higher pitched notes after listening to examples of the middle of each range (i.e., D2 for low and D3 for high). Based on this indicated preference, participants heard either the lower range of pitches to imitate (A2 to A3) or the higher range of pitches to imitate (A3 to A4). Participants then imitated three types of

[11] Pfordresher et al. (2010) administered 38 total musical stimuli with intervals ranging from unisons to octave jumps; however, to keep testing time to a minimum, only stimuli outlining unisons to perfect fifths were included here.

sequences: note, interval, and melody sequences. In note sequences, the same pitch was repeated five times. In the interval sequences, the pitch changed between the second and third note (i.e., both descending and ascending intervals up to a perfect fifth). Finally, in the melody sequences, all five pitches alternated in various patterns and pre-chosen intervals of a third, fourth, and fifth were analyzed for accuracy. Participants' analyzed according to the same as above.

In this task, multiple participants sang in a different octave than the musical stimuli they selected to hear (e.g., a participant heard A3 but sang A2), presumably because they realized the stimuli they heard were not within their vocal range. Because of this and the perceptual similarity between pitch chromas of different octaves (i.e., octave equivalence, see e.g., Deutsch, 2013), pitches from different octaves (ranging from A1-E4) were counted as accurate. To analyze note sequences, the utterance containing five notes was extracted using Audacity (v 1.3.12, Audacity Team, 2010). Then, Praat (v 6.0.19, Boersma & Weenink, 2016) and PraatR (v 2.4, Albin, 2014) were used to find the mean pitch in Hz. In order to account for participants singing in either octave, the absolute distance in cents was calculated from sung note (e.g., ~A2) to the intended note in multiple octaves (e.g., A1, A2, A3), and the smallest distance was used as the accuracy score. To analyze the interval and melody sequences, the two "/da/"s containing the notes of interest were extracted using Audacity (v 1.3.12, Audacity Team, 2010). Then, Praat (v 6.0.19, Boersma & Weenink, 2016) and PraatR (v 2.4, Albin, 2014) were used to find the mean pitch in Hz of each of the two utterances. Again, to account for participants singing in different octaves, the absolute distance in cents was calculated from each sung note to the intended note in multiple octaves, and the smallest distance was used as the individual note accuracy score. Analyses for the interval and melody sequences here differ from those used by Pfordresher et al. (2010): instead of calculating the interval accuracy (by calculating the distance between the two pitches and disregarding whether the sung pitch matched the target pitch), the individual note accuracy was used to indicate performance for these conditions. Scores for each participant were averaged accuracy for all note, interval, and melody sequences.

Timing Perception Measures. These four tasks – PROMS-S Rhythm, PROMS-S Tempo, BAT Perception, and Duration Discrimination Threshold – all required perceptual judgments of aspects related to timing.

PROMS-S. To measure timing perception, the PROMS-S Rhythm and Tempo subtests (Law & Zentner, 2012; Zentner & Strauss, 2017) were administered. These shortened subtests followed the same presentation (i.e., participants heard the first stimulus twice, then heard the comparison stimulus) and answer procedures (i.e., answering “definitely different,” “probably different,” “I don’t know,” “probably same,” or “definitely same”) as the subtests outlined in the Pitch section above. Also as described above, participants’ raw accuracy scores were weighted by confidence and were transformed to d’ for both the Rhythm and Tempo subtests.

The PROMS-S Rhythm subtest contained sequences of varying length (seven to twelve notes) voiced with a percussive “rim shot” timbre. The eight trials ranged in difficulty, with easier trials consisting of simple rhythms of mostly quarter notes and eighth notes, and more complex trials containing complex rhythm structures with mostly eight and sixteenth notes (see Figure 3 for an example of a different trial).

Participants' raw scores were calculated to include confidence, and were transformed to d'.

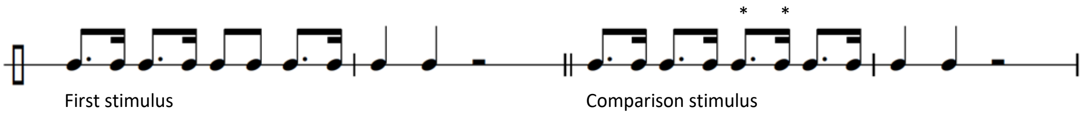


Figure 3. Example of a different stimulus from the Rhythm subtest (from Law, 2012). The asterisk indicates the changed rhythm within the second comparison stimulus.

For the PROMS-S Tempo subtest, participants heard two musical stimuli that were either the same or different in speed. Three different musical stimuli were composed and used[12]: a rim shot voice similar to the Rhythm subtest (with one layer of sound), a conga and shaker (with two layers of sound), and drums, bass, harmony, and melody (with multiple layers of sound). The eight trials ranged in difficulty, with easier trials containing larger tempo changes, and more complex trials containing smaller tempo changes. Participants' raw scores were calculated to include confidence, and were transformed to d'.

Timing Threshold Discrimination. In order to measure perceptual acuity of timing, participants completed the Duration Discrimination Test from the Psychoacoustics Toolbox in Matlab (Grassi & Sorenzo, 2009; Sorenzo & Grassi, 2014). In this adaptive task, participants completed 2 blocks of 30 trials each. In each trial, participants heard three serially presented pure tones (3AFC), and determined which was the longest in duration. In this 3AFC task, two of the tones had the same duration and the other was presented at a variable duration (either below or above the

[12] Three varying stimuli were used instead of only one stimulus because participants reported that it was difficult to remember the various tempi when only hearing one song multiple times (Law, 2012).

participants discrimination threshold). Based on each participants' performance, the threshold was calculated using the maximum likelihood estimation procedure, as described above in the Pitch Threshold Discrimination task. Again, each participant's discrimination threshold was calculated in both blocks, and the average of the two blocks was used as the final timing threshold discrimination score.

Beat Alignment Test - Beat Perception. In order to measure rhythm and timing perception, a shortened version (Müllensiefen et al., 2014) of the perception subset (Beat Perception in a Musical Passage) of the Beat Alignment Test (BAT, Iverson & Patel, 2008) was administered in PsychoPy (v 1.82.0, Peirce, 2007; 2009). (Note that the BAT also contains multiple timing production tasks, which are outlined in the following section.) In this task, participants listened to 17 musical excerpts[13] with metronome beeps superimposed, and determined whether or not the superimposed metronome beeps synchronized with the musical passage. The stimuli consisted of twelve different instrumental songs from three different genres (i.e., rock, jazz, pop orchestral). Beeps were either on the beat, out of phase (i.e., consistently ahead or behind the beat of the music), or played at a different tempo (i.e., were faster or slower than the music). For the stimuli that did not have metronome beeps on the beat, the metronome beep tempo was shifted by either 2, 10, or 17.5%. Participants received three practice trials for exposure to all three conditions before completing the 17 trials. Participants' scores were calculated as proportion accuracy.

Timing Production Measures. In these three tasks – Synchronization to a Metronome, Synchronization Continuation, and Song Synchronization – participants'

[13] In order to minimize testing time, only a subset of the original 24 trials (in the BAT) were presented (following version 1.0 of the task in Müllensiefen et al., 2014).

timing production was measured by consecutive tapping on the spacebar of a keyboard.

Synchronization to a Metronome & Synchronization Continuation. In the first timing production task, two measures were collected within the same task (i.e., synchronization to a metronome and synchronization continuation), and are described together here. To measure timing production, a modified version of the Beat Alignment Test (BAT; Iverson & Patel, 2008) was administered in PsychoPy (v 1.82.0, Peirce, 2007; 2009). The original BAT contains two production subtests: synchronization to a metronome and synchronization to musical passages. In the first synchronization to a metronome portion, participants used their dominant hand to tap along on the space bar to a 30 second clip[14] of a metronome beep at various tempi (i.e., inter-onset interval (IOI) of 400 ms, 550 ms, and 700 ms, taken from Iverson & Patel, 2008). Immediately following the 30 seconds of metronome beeps, participants were instructed to continue tapping in silence and to try to maintain the same beat (at the same speed) for an additional 10 seconds (i.e., synchronization continuation). This synchronization continuation measure tested whether participants were able to maintain the same beat in silence on their own (following Dalla Bella et al., 2017). One practice trial (i.e., 10 second clip of a metronome beeping at 530 IOI) was given to allow participants to familiarize themselves with the task of tapping along to the metronome, then continuing to tap in silence afterwards.

[14] In the pre-registration, it was proposed to have participants tap to a metronome beep for 20 seconds, then tap in silence for 20 seconds. However, due to a programming error, the full 30 second metronome beep clip from the BAT was used (Iverson & Patel, 2008), and participants tapped in silence for the following 10 seconds.

Although participants completed these two measures within the same task, two timing synchronization accuracy scores were calculated for tapping along with the metronome and the continuation of tapping without the metronome. First, the first two and last two taps for each trial were discarded to account for start-up and wind-down effects (following Fujii & Schlaug, 2013) and to ensure that taps in silence did not overlap with the taps to the metronome during analysis. Inter-tap intervals (ITIs) were calculated as the successive difference between the remaining tap times. To remove artifacts, all ITIs below 100ms were removed as well as any outliers, defined as any ITI greater than 3 times the interquartile range (IQR) of each trial (following Dalla Bella et al., 2017; Jakubowski, Bashir, Farrugia, & Stewart, 2018). To calculate timing production variability, the coefficient of variation (CV) was calculated. The CV is a normalized measure of tapping variability calculated by dividing the standard deviation of the ITI for each trial by the mean of the ITI for each trial.

Song Synchronization. A modified version of the Synchronization to Musical Passages subtest (from the BAT, Iverson & Patel, 2008) was administered from the Harvard Beat Assessment Test[15] (H-BAT, Fujii & Schlaug, 2013) in PsychoPy (v 1.82.0, Peirce, 2007; 2009). This subtest from the H-BAT included three musical stimuli taken from the BAT (i.e., “Hurts So Good” by John Mellencamp, “Tuxedo Junction” by Glenn Miller, and “A Chorus Line” by the Boston Pops), which were each presented three times in three different tempi. In this subtest, participants were

[15] The H-BAT version was used instead of the original 12 different stimuli from the BAT because these three songs yielded the highest pulse clarity measures (assumed to have the least ambiguous pulse), were more controlled than the original stimuli, and in order to keep testing time to a minimum.

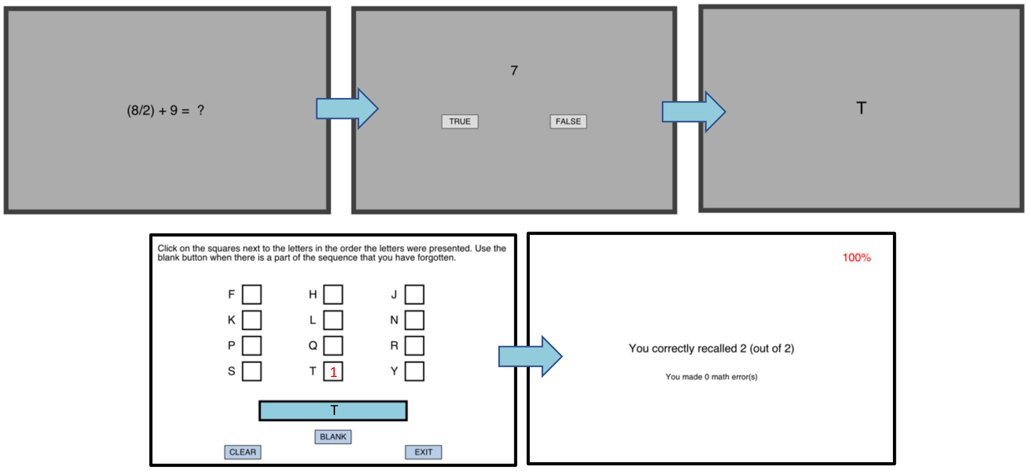
again instructed to use their dominant hand to tap as closely to the beat of the song as possible. The same trimming criteria (i.e., eliminate first two and last two taps and discard any ITIs below 100 ms and above 3*IQR) as above were used, and participants' scores were again calculated as the coefficient of variation and averaged across trials.

Ancillary Measures

In addition to the primary tasks described above, measures assessing several other factors that are likely relevant to musical ability were assessed (i.e., musical training, working memory, intelligence, personality, and socio-economic status).

Musical Training. To assess participants' musical training, the Goldsmith Musical Sophistication Index questionnaire was administered (Gold-MSI, Müllensiefen et al., 2014). This self-report questionnaire measures "musical sophistication" (defined as a "construct that can refer to musical skills, expertise, achievements, and related behaviours across a range of facets"; Müllensiefen et al., 2014, p. 2) with questions in five subscales: musical training, active engagement, perceptual abilities, singing abilities, and emotions. Measures were collected from all subscales, however only the musical training subscale was used in the current analysis. The musical training subscale differs from previous ways used to measure musical experience in that it contains seven questions regarding musical training, which include: years of instrument training, years of music theory training, regular daily practice, the number of hours practiced at peak of interest, the number of instruments played, whether compliments about music performances have been received, and whether they consider themself a musician. Since this measure takes into account how long one has taken music lessons as well as the intensity of practice, participants' scores from the musical training subtest provided a continuous and more robust measure of musical training (rather than only looking at duration of music lessons like most studies). Participants answered these questions on 7-point Likert scales. Two items were reverse scored, and participants' music training scores were calculated as the sum of the seven music training subscale items.

Working Memory. To examine the relationship between general WM ability and musical ability, a shortened version of the automated operation span (AO-Span, Unsworth, Heitz, Schrock, & Engle, 2005) was administered. In this task, participants first completed three practice blocks. The first practice block served to allow participants to familiarize themselves with the math portion of the task. Here, a math problem was shown on the screen (e.g., (1*4) + 2 = ?), and participants clicked as soon as they solved the problem. Next, on the following screen, participants saw a number (e.g., 6) and clicked "true" or "false," and were given feedback. For each participant, the mean reaction time to answer all 15 practice math problems was calculated and the mean plus 2.5 *SD* was used as the math time limit during the last practice block and the real trials. This accounted for individual differences in solving the math problems and ensured that participants were dual tasked. The second practice block allowed participants to familiarize themselves with the letter span portion of the task. Here, participants saw letters presented serially (only letters F, H, J, K, L, N, P, Q, R, S, T, and Y were used), then were presented with a test screen containing all of the letters. Participants were instructed to click on the boxes next to the letters they recalled in order, then received feedback. Participants completed three practice trials with letter set sizes of two. The third practice block combined the first math and letter tasks: first participants solved a math problem, saw a letter, solved another math problem, then saw another letter. They repeated this process until they recalled all the letters seen at the very end of the trial, after which they received feedback on their math performance and letter span performance. Participants were told they would only have a limited amount of time to solve the math problem, and if they reached the time limit without clicking the mouse, that math problem was counted as incorrect. Additionally, they were told to not let their math performance fall below 85%. Participants completed three practice trials all with a set size of 3 (see Figure 4 for a schematic of the test).



Click on the squares next to the letters in the order the letters were presented. Use the blank button when there is a part of the sequence that you have forgotten.

You correctly recalled 2 (out of 2)

You made 0 math error(s)

Figure 4. Test schematic of the automated operation span (adapted from Unsworth et al., 2005).

For the real trials, participants completed eight trials with two set sizes of four, two set sizes of five, two set sizes of six, and two set sizes of seven. After each trial, feedback on the math performance and letter span was provided. The WM score was calculated as the sum of the total number of correctly completed set sizes (i.e. absolute span), which ranged from 0-44.

Intelligence. To investigate the relationship between general fluid intelligence and musical ability, a shortened version[16] of the Raven's Advanced Progressive Matrices (RAPM, Raven, Raven, & Court, 1991) was administered. In each trial, participants saw a 3x3 matrix of various geometric shapes, and the bottom right cell in the matrix was always blank. Participants were instructed to look at the patterns both across and down the matrix, then select which of the eight possible answer choices best completed the pattern. Participants were then told to complete as many of the 18 trials as possible within 10 minutes (following Kane et al., 2004; Kane et al., 2007). Participants' scores were calculated as the proportion of correct answers[17].

Personality. To assess the relationship between Openness to Experience and musical ability, the Big 5 inventory[18] (John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008) was administered. In this measure, participants used 5-point Likert scales (i.e., "Disagree Strongly," "Disagree a little," "Neither agree nor disagree," "Agree a little," or "Agree strongly") to answer questions regarding a number of personal characteristics. There were 10 questions regarding Openness to Experience (e.g., "I am someone who is original, comes up with new ideas", "I am

[16] The original RAPM Set 2 contained 36 items, and only the 18 odd trials were used here to shorten testing time (following Kane et al., 2004; Kane et al., 2007).

[17] Accuracy was calculated as number of correct answers divided by 18 even though not all participants completed all 18 trials. Mean number of trials completed was 17.15, range 11-18.

[18] Although the main variable of interest here was openness to experience, the full Big 5 inventory was administered to maintain its psychometric properties.

someone who values artistic, aesthetic experiences", "I am someone who is sophisticated in art, amusic, or literature"). Two items were reverse scored, and participants' Openness scores were calculated as the average response to the 10 Openness items.

Socioeconomic Status. To measure SES, the MacArthur Scale of Subjective Social Status was administered (Adler & Stewart, 2007). Here, participants indicated where they believe they stood (in terms of money, education, and job status) relative to others in the U.S., on a scale of 1 to 10[19]. This subjective measure of SES has been shown to better predict outcomes (e.g., health status) than objectives measures of SES such as income or occupational status (Singh-Manoux, Marmot, & Adler, 2005).

Procedure

Participants completed this study in either two sessions (online and in-person) or in one session (in-person only)[20]. In the two session format, the first portion of the

[19] Participants were also asked to self-report their parents' income and education levels (on 9-point and 6-point scales, respectively) as measures of objective SES. However, multiple participants indicated that they did not know or did not wish to provide this information (specifically, 41 participants did not report their father's income, 39 participants did not report their mother's income, and one did not report their father's education level). Reassuringly, for those participants who did provide these ratings, the subjective measure of SES correlated significantly with both parental education ($r(162) = .21, p = .007$) and parental income ($r(113) = .41, p < .001$).

[20] Testing changed to take place in only one session for 2 reasons: 1) the total testing time ended up being shorter than anticipated, making it feasible to combine both sessions, and 2) because the first online portion of the study was worth .5 SONA credits, many participants were unnecessarily excluded from participation if they had already completed the maximum amount of *online* studies for a class (2 SONA credits). 25 participants completed the one-session study and the remaining participants completed the two-session study.

study was administered remotely on Qualtrics online, and participants answered questionnaires about their background (i.e., hearing screen, Gold-MSI, SES, personality measures). During the second in-person session, participants completed the pitch tasks (perception and production), the rhythm tasks (perception and production), and the WM and IQ tasks, which lasted about one hour and 45 minutes to two hours. For the one session format, participants completed all tasks in the same order, but filled out the online Qualtrics questionnaires on a testing computer in the lab right before performing the other tasks.

To ensure test order did not influence individual differences results, participants completed the tasks in the same fixed order. First, participants completed the five PROMS-S tasks (Melody, Rhythm, Tuning, Tempo, Pitch), then they completed the two threshold discrimination tasks (Pitch and Timing). Then, to minimize fatigue, participants took a mandatory 3-5 minute break during which they were encouraged to stand and stretch. After the first break, participants warmed up and completed the pitch imitation tasks (Familiar Songs, Mowrer Test of Tonal Memory, Melody Imitation), Chord Analysis, and the BAT Perception task. Next, participants took a second mandatory 3-5 minute break before completing the Timing production tasks (Synchronization to a Metronome, Synchronization Continuation, and Song Synchronization), WM task, and the IQ test. This order was chosen so that the pitch and timing tasks alternated when possible, the three blocks of the study were around the same duration, and to maximize efficiency (i.e., the PROMS tasks were completed together because they were presented in the same experimental platform).

Table 2.
List of Musical Tasks

	Pitch	Timing
Perception	PROMS-S - Pitch	PROMS-S - Rhythm
	PROMS-S - Melody	PROMS-S - Tempo
	PROMS-S - Tuning	BAT - Beat Perception
	Chord Analysis	Timing Threshold Discrimination
	Pitch Threshold Discrimination	
Production	Familiar Song Imitation	BAT - Synchronization to a Metronome
	Mowrer Test of Tonal Memory	Synchronization Continuation
	Melody Imitation (note, interval, melodies)	H-BAT - Synchronization to Musical Passage

Chapter 3: Results

Descriptive statistics for participant demographics, ancillary measures, and musical measures are provided in Tables 3 and 4. Cronbach's alpha (α) was calculated for internal consistency and is reported where applicable (Cronbach, 1951). However, since Cronbach's alpha estimates the lower bound of reliability and assumes tau-equivalence for all test items (i.e., test items all measure the same construct with the same degree of precision and amount of error), McDonald's omega (ω) was also calculated where possible and reported as a better estimate of internal consistency (McDonald, 1999; see Revelle & Zinbarg, 2009 for a discussion of both alpha and omega). McDonald's omega uses factor analysis to estimate the extent to which test items are capturing a construct without the assumption of tau-equivalence (Revelle, 2018). Reliabilities found for the tasks used here are similar to those reported in previous studies. The reliabilities for the PROMS-S tasks are similar to those in Zentner and Strauss (2017), although lower than the reliabilities found in the original version of the PROMS (Law & Zentner, 2012). Threshold discrimination tasks show comparable reliabilities to those reported in Kidd et al (2007), and the musical training and working memory reliabilities are also similar to those reported previously (Müllensiefen et al., 2014; Unsworth et al., 2005).

Table 3
Descriptive Statistics for Participant Demographics and Ancillary Measures

Measure	Mean	SD	Min	Max	Skewness	Kurtosis	α	ω
Age (Years)	19.85	1.69	18.00	29.00	1.99	7.01	-	-
SES (MacArthur Ladder)	6.42	1.38	3.00	10.00	-0.08	-0.12	-	-
Openness to Experience (BIG 5)	5.24	0.57	3.50	6.50	-0.11	-0.29	0.80	0.85
WM (Absolute Span)	18.36	11.36	0.00	44.00	0.34	-0.61	0.69	0.76
IQ (RAPM Accuracy)	0.54	0.18	0.06	0.89	-0.18	-0.37	0.75	0.78
Musical Training (Gold-MSI)	23.89	8.91	7.00	43.00	-0.05	-0.75	0.84	0.89

Note. Total *N* = 165

Table 4

Descriptive Statistics for Musical Measures

Measure	Mean	SD	Min	Max	Skewness	Kurtosis	α	ω
PROMS-S Melody (d')	1.15	0.87	-0.80	3.29	-0.02	-0.48	0.61	0.67
PROMS-S Tuning (d')	1.13	0.69	-0.48	3.07	0.36	0.08	0.65	0.73
PROMS-S Pitch (d')	0.61	0.69	-1.47	2.21	0.16	-0.56	0.46	0.73
PROMS-S Rhythm (d')	1.31	0.92	-1.53	3.07	-0.04	-0.25	0.52	0.62
PROMS-S Tempo (d')	1.52	0.78	-0.38	3.07	-0.04	-0.53	0.49	0.53
Pitch Discrimination Threshold (change in Hz)	8.40	10.66	0.76	60.76	2.88	8.65	0.85	-
Duration Discrimination Threshold (change in ms)	54.15	24.49	15.07	166.27	1.78	4.56	0.75	-
Familiar Songs (cents)	290.35	223.89	15.80	1348.47	1.43	2.76	0.64	0.96
Mowrer Test (accuracy)	2.31	1.92	0.00	6.00	0.40	-1.12	0.77	0.87
Melody Imitation (cents)	243.58	99.46	125.57	732.61	1.91	5.62	0.80	0.91
Chord Analysis (accuracy)	0.52	0.14	0.15	0.90	0.14	-0.21	0.53	0.60
BAT Perception (accuracy)	0.70	0.15	0.35	1.00	0.01	-0.61	0.56	0.62
Metronome Synchronization (CV)	0.05	0.01	0.03	0.12	1.32	2.79	0.75	-
Synchronization Continuation (CV)	0.05	0.02	0.02	0.16	2.40	10.61	0.62	-
Song Synchronization (CV)	0.09	0.08	0.04	0.44	2.62	6.70	0.93	0.96

Note. Total *N* = 165, except for Pitch Discrimination *N* = 161, Duration Discrimination *N* = 153, Familiar Songs *N* = 145, Mowrer Test *N* = 163, Melody Imitation *N* = 163, Chord Analysis *N* = 164, Metronome Synchronization *N* = 164, Synchronization Continuation *N* = 163

McDonald's omega (ω) calculated where possible. For some measures (i.e., both Threshold Discrimination tasks, Metronome Synchronization task, and Synchronization Continuation tasks), ω could not be calculated due to the low number of response outcomes.

Tables 5 and 6 show correlations between each of the measures. To facilitate interpretation, the scores for each task were adjusted so that higher scores would indicate better performance (specifically, both Threshold Discrimination scores, all three Timing Production task scores, the Familiar Song Imitation score, and the Melody Imitation score were multiplied by -1). Here, performance on many of the musical tasks have positive correlations with one another. Importantly, measures of each intended construct showed strong correlations (except for the relationship between the PROMS-S Pitch subtest and Chord Analysis task), and the magnitude of the correlations suggest that these tasks are tapping related abilities, but are not completely redundant measures (i.e., high correlations). Another approach to calculating correlations is to correct for the attenuation of tasks with lower reliabilities (Spearman, 1904). These disattenuated correlations (i.e., an estimate of the true correlation of two variables if both variables were perfectly reliable and with no measurement error) also show the same pattern of results with mostly stronger relationships and are reported in Appendix A and Appendix B. These disattenutated correlations may more closely represent the relationships represented between tasks in the more comprehensive latent variable framework described below.

Table 5

Correlation Matrix of Musical Measures

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Pitch Perception	1. PROMS-S Melody														
	2. PROMS-S Tuning	0.32***													
	3. PROMS-S Pitch	0.27***	0.27***												
	4. Pitch Threshold	0.23**	0.23**	0.22**											
	5. Chord Analysis	0.23**	0.28***	0.11	0.35***										
Pitch Production	6. Familiar Songs	0.17*	-0.03	0.08	0.18*	0.11									
	7. Mowrer Test	0.51***	0.36***	0.37***	0.24**	0.33***	0.19*								
	8. Melody Imitation	0.30***	0.13	0.20*	0.06	0.24**	0.32***	0.53***							
Timing Perception	9. PROMS-S Rhythm	0.29***	0.26***	0.19*	0.16*	0.14	0.13	0.28***	0.06						
	10. PROMS-S Tempo	0.16*	0.24**	0.13	0.17*	0.05	-0.03	0.26***	0.07	0.20*					
	11. Duration Threshold	0.20*	0.23**	0.19*	0.46***	0.26**	-0.02	0.14	0.03	0.17*	0.26**				
	12. BAT Perception	0.29***	0.24**	0.26***	0.28***	0.27***	0.06	0.41***	0.17*	0.18*	0.31***	0.39***			
Timing Production	13. Metronome Sync	0.34***	0.21**	0.20*	0.49***	0.31***	0.17*	0.31***	0.09	0.21**	0.27***	0.33***	0.36***		
	14. Sync Cont	0.10	0.09	0.11	0.34***	0.23**	0.14	0.20**	0.11	0.12	0.17*	0.22**	0.29***	0.50***	
	15. Song Sync	0.19*	0.27***	0.14	0.16*	0.25**	-0.03	0.28***	0.07	0.25**	0.27***	0.20*	0.33***	0.28***	0.29***

Note. *$p < .05$, **$p < .01$, ***$p < .001$

Table 6

Correlation Matrix of Ancillary Measures and Musical Measures

	1	2	3	4	5
1. Music Training					
2. Openness	0.23**				
3. IQ	0.16*	0.12			
4. SES Ladder	-0.07	0.05	0.02		
5. WM	0.14	0.12	0.25**	0.09	
PROMS-S Melody	0.43***	0.15	0.16*	-0.05	0.13
PROMS-S Tuning	0.39***	0.03	0.23**	0.06	0.17*
PROMS-S Pitch	0.23**	0.06	0.09	-0.04	0.07
Pitch Threshold	0.25**	-0.01	0.28***	0.02	0.13
Chord Analysis	0.32***	0.04	0.26***	0.12	0.13
Familiar Songs	0.24**	0.03	0.16*	-0.08	0.15
Mowrer Test	0.47***	0.16*	0.04	0.01	0.19*
Melody Imitation	0.33***	-0.03	0.02	-0.07	0.09
PROMS-S Rhythm	0.30***	0.18*	0.10	-0.05	0.12
PROMS-S Tempo	0.18*	0.02	0.12	-0.11	0.09
Duration Threshold	0.15	0.02	0.35***	0.03	0.02
BAT Perception	0.40***	0.04	0.27***	0.09	0.16*
Metronome Sync	0.34***	0.04	0.14	0.06	0.06
Sync Cont	0.17*	-0.10	0.16*	0.04	0.08
Song Sync	0.34***	0.06	0.22**	-0.12	0.18*

Note. **p* < .05, ***p* < .01, ****p* < .001

Structural Analyses/Model Comparisons

In order to improve upon the "task impurity problem" (i.e., even if a complex task does measure the construct of interest, it also taps other types of processes as well), the analyses relied on latent variable analysis to obtain estimates of the latent ability underlying performance on a set of theoretically related tasks (e.g., Miyake et al., 2000; Okada & Slevc, 2018). That is, because a single task measuring a construct of interest may not be indicative of someone's true score (e.g., could include measurement error), multiple measures of each construct of interest were administered. By using latent variable analysis, one can estimate what is common between the tasks measuring a given construct and get a better estimate of the underlying component of interest removed from task-specific effects.

In order to answer the research questions below, a series of confirmatory factor analyses (CFA) were assessed to determine the best model of musical ability that fit the data. All analyses were conducted with the R Statistical Platform (v. 3.4.1, R Core Team, 2017) using the package lavaan (v. 0.5-23.1097, Rosseel, 2012). Each model was run with robust maximum likelihood estimation (MLR), which provided robust standard errors and a Yuan-Bentler scaled chi-square test statistic that is robust to non-normality (Yuan & Bentler, 2000). Full Information Maximum Likelihood (FIML) was used for missing data in order to obtain estimates with partial data (as opposed to pair-wise or list-wise deletion, Beaujean, 2014).

For each model, the following model fit indices are reported: Chi-square test of model fit (χ^2), Comparative Fix index (CFI), Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Square Residual (SRMR), and the Akaike Information Criterion (AIC). The Chi-square Test of Model Fit tests the null hypothesis that the observed data are no different from the expected population covariance matrix from the model (i.e., that the model fits the data). The alternative hypothesis is that our observed data do not fit the population covariance specified by our model. Thus, a non-significant χ^2 means that the model fits the data well. However, this statistic is influenced by sample size, so it is reported alongside other model fit indices. The CFI is classified as a comparative index of model fit because it indicates “improvement in model fit by comparing the hypothesized model in which structure is imposed with the less restricted nested baseline model” (Byrne, 2013, p. 72). A CFI above .95 is considered good model fit (Hu & Bentler, 1999). The RMSEA and SRMR are classified as absolute indices of model fit because they do not compare the hypothesized model with a "reference model in determining the extent of model improvement; rather, they depend only on determining how well the hypothesized model fits the sample data" (Byrne, 2013, p. 72). A RMSEA less than .05 and a SRMR under .08 show good model fit (Hu & Bentler, 1999). The AIC is a measure of goodness of fit, and it penalizes for the addition of more parameters; a smaller AIC is indicative of better model fit. If the two models being compared were nested, a Satorra-Bentler scaled χ^2 difference test was conducted to determine if one model fit significantly better than the other (Satorra & Bentler, 2001). However, if the two models being compared were non-nested, the model with the lowest AIC was deemed the better fitting model.

In the following figures, all observed, measured variables are represented in squares and unobserved, latent factors are represented in circles. The single headed arrows from latent factors to measured variables are standardized factor loadings. Double headed arrows between latent factors represent the correlation between the two factors – squaring this value gives shared variance between the factors. For path models, single headed arrows from a measured task or latent variable to another latent variable can be interpreted as regression coefficients (e.g., a one unit increase in WM equals on average, a one unit increase in musical ability (Beaujean, 2014).

Is Musical Ability a Unitary, General Construct, or Are Pitch and Timing Dissociable Abilities?

In order to answer this question, two models (see Figure 5 below) were fitted and compared using the lavaan package (v 0.5.23.1097, Rosseel, 2012) in R (v. 3.4.1, R Core Team, 2017). The first model (Model 1) was composed of a unitary Musical Ability factor, with all 15 musical tasks loading onto it. This model showed acceptable model fit (see Table 7 for all model fit indices), and all tasks loaded significantly onto the unitary Musical Ability factor. The second model (Model 2) was a two-factor model with separate, but correlated Pitch and Timing factors. The Pitch factor items consisted of the Pitch, Melody, and Tuning subtests of the PROMS-S, Chord Analysis, Pitch Discrimination Threshold, and all pitch production measures. The Timing factor items consisted of the Rhythm and Tempo subtests of the PROMS-S, Duration Discrimination Threshold, BAT Perception subtest, and all timing production measures[21]. Model 2 also showed acceptable model fit (see Table 7), and again all tasks loaded significantly onto their respective latent factors.

In order to determine whether musical ability is a unitary construct or if it is more appropriate to have separate Pitch and Timing factors, a χ^2 difference test was run, and Model 2 showed significantly better fit (Satorra-Bentler scaled χ^2 difference (1, N = 165) = 22.33, $p < .001$). In Model 2, the correlation between the latent factors of Pitch and Timing was $r = .72$, $p < .001$.

[21] The OSF pre-registration erroneously stated that the BAT Perception subtest should be included in the Pitch factor; however, it is included here in the Timing factor.

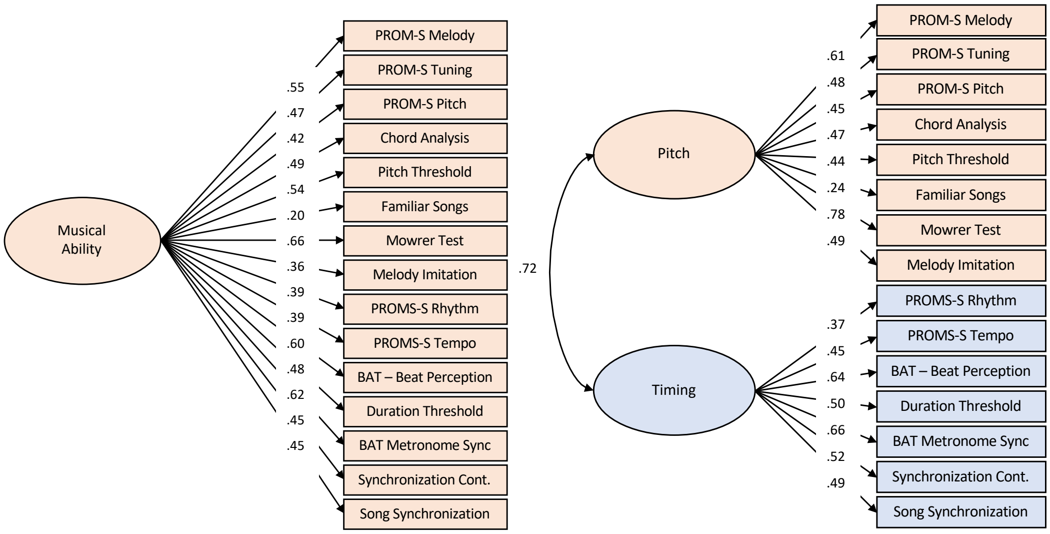


Figure 5. 1-factor model (Model 1) where all tasks load onto a general musical ability factor, and a 2-factor model (Model 2) with Pitch and Timing factors

Table 7
Model Fit Indices

Model	x2	df	AIC	SRMR	RMSEA	CFI
1	216.321	90	6583.014	0.078	0.088	0.756
2	188.640	89	6559.020	0.078	0.079	0.808
4A	95.931	73	6500.109	0.052	0.040	0.959
4B	156.067	84	6534.697	0.068	0.068	0.865

Note. Model 4A, with the best fit indices, is highlighted in blue.

Are Perception and Production Abilities Dissociable?

Given that the two-factor model with Pitch and Timing factors fit significantly better than the one-factor model above, two different four-factor models incorporating Pitch, Timing, Perception, and Production factors were fit.

Model 4A contained the following four factors: Pitch, Timing, Perception, and Production (see Figure 6). The Pitch and Timing factors contained the same items as in Model 2. The Perception factor items included: all PROMS-S subtests, Chord Analysis, Pitch and Duration Thresholds, and BAT Beat Perception. The Production factor items included all three singing measures and all three tapping measures. This model showed good model fit, and showed better fit than model 2 (i.e., lower AIC, see Table 7).

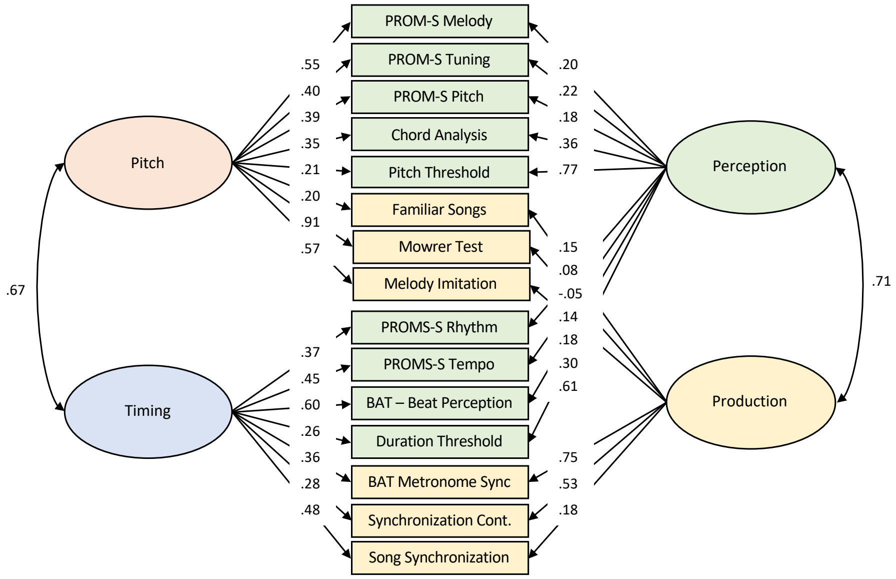


Figure 6. Model 4A with Pitch, Timing, Perception, and Production factors

Model 4B contained the following four factors: Pitch Perception, Pitch Production, Timing Perception, and Timing Production (see Figure 7). The Pitch Perception factor contained the Pitch, Melody, and Tuning subtests of the PROMS-S, Chord Analysis, and Pitch Discrimination Threshold. The Timing Perception factor contained the Rhythm and Tempo subtests from the PROMS-S, Duration Discrimination Threshold, and BAT Perception subtest. The Pitch Production factor included the Familiar Songs Imitation task, Mowrer Test of Tonal Memory, and Melody Imitation task. Lastly, the Timing Production factor included the Metronome Synchronization task, Synchronization Continuation, and Song Synchronization. Model 4B showed acceptable model fit, but did not fit as well as Model 4A (i.e., Model 4B had higher AIC than Model 4A, see Table 7).

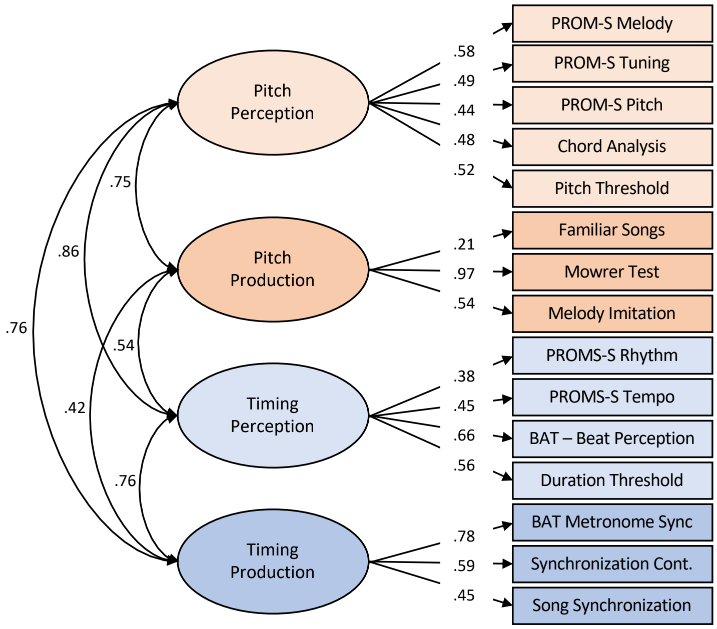


Figure 7. Model 4B with Pitch Perception, Pitch Production, Timing Perception, and Timing Production factors

With Model 4A having the overall best fit, the relationships between pitch and timing and perception and production were evaluated. In this final model, the Pitch and Timing factors were significantly correlated ($r = .67$, $p < .001$) and the Perception and Production factors were also significantly correlated ($r = .71$, $p < .001$)). These correlations suggest related, but separate processes underlying performance on these tasks.

In order to assess the relationships between finer-grained auditory acuity and the processing of more complex music, factor loadings and correlations between the Threshold Discrimination measures and more complex musical tasks were assessed. Both of the Threshold Discrimination tasks were strongly and positively correlated with one another, and they both loaded significantly onto their respective factors (within Model 4A). This suggests that performance on these two tasks was influenced by a similar latent process underlying both fine-grained auditory processing and more complex musical processing (e.g., the Pitch factor influences both Pitch Threshold Discrimination performance and more complex musical tasks, such as the PROMS-S Melody subtest).

How Does Musical Ability Relate to the Ancillary Measures?

The goal of this portion of the study was to determine musical ability's relationship with other relevant variables. Because most of the hypothesized relationships between musical ability and the ancillary measures were not specific to sub-aspects of musical ability (and because our one factor model showed acceptable model fit), we proceeded to determine these relationships by fitting the one-factor model of Musical Ability with separate path models for each ancillary ability of interest (i.e., musical training, WM, IQ, Openness, SES). By examining the significance of the path coefficient with the one-factor model, the relationship between these variables and overall musical ability was assessed. Although the primary analyses here rely on a one factor model of musical ability, given the best fitting four factor model above, a set of exploratory analyses were performed examining the relationships between each of the four factors and ancillary measures to investigate the potential for more specific relationships.

Musical training, as measured by the seven items on the Gold-MSI, did predict overall musical ability (*estimate* = .70, *p* < .001). With the four factor model of musical ability, musical training predicted latent Pitch (*estimate* = .63, p < .001) and Timing (*estimate* = .65, *p* = .004) abilities, but neither Perception (*estimate* = .21, *p* = .29) nor Production (*estimate* = .20, *p* = .28).

Working Memory, as measured by the Operation span, did predict overall musical ability (*estimate* = .25, *p* = .005). Just as with musical training, within the four factor model of musical ability, working memory predicted latent Pitch (*estimate* = .21, *p* = .026) and Timing (*estimate* = .24, *p* = .035) abilities, but not Perception (*estimate* = .082, *p* = .29) nor Production (*estimate* = -.013, *p* = .90) abilities.

IQ, as measured by the Raven's Advanced Progressive Matrices, did predict overall musical ability (*estimate* = .37, *p* = .001). With the four factor model of musical ability, IQ predicted latent Timing (*estimate* = .26, *p* = .042) and Perception (*estimate* = .45, *p* < .001) abilities, but neither Pitch (*estimate* = .040, *p* = .67) nor Production (*estimate* = .088, *p* = .45) abilities.

Openness, as measured by the ten items on the BIG5, was not predictive of overall musical ability (*estimate* = .092, *p* = .35), nor any of the latent factors within the four factor model of musical ability: neither Pitch (*estimate* = .16, *p* = .097), Timing (*estimate* = .11, *p* = .39), Perception (*estimate* = -.048, *p* = .60), nor Production (*estimate* = -.031, *p* = .78).

SES, as measured by the MacArthur ladder did not predict overall musical ability (*estimate* = .016, *p* = .84), nor any of the latent factors within the four factor model of musical ability: neither latent Pitch (*estimate* = -.014, *p* = .87), Timing (*estimate* = -.092, *p* = .52), Perception (*estimate* = .096, *p* = .35), nor Production (*estimate* = .11, *p* = .30) abilities.

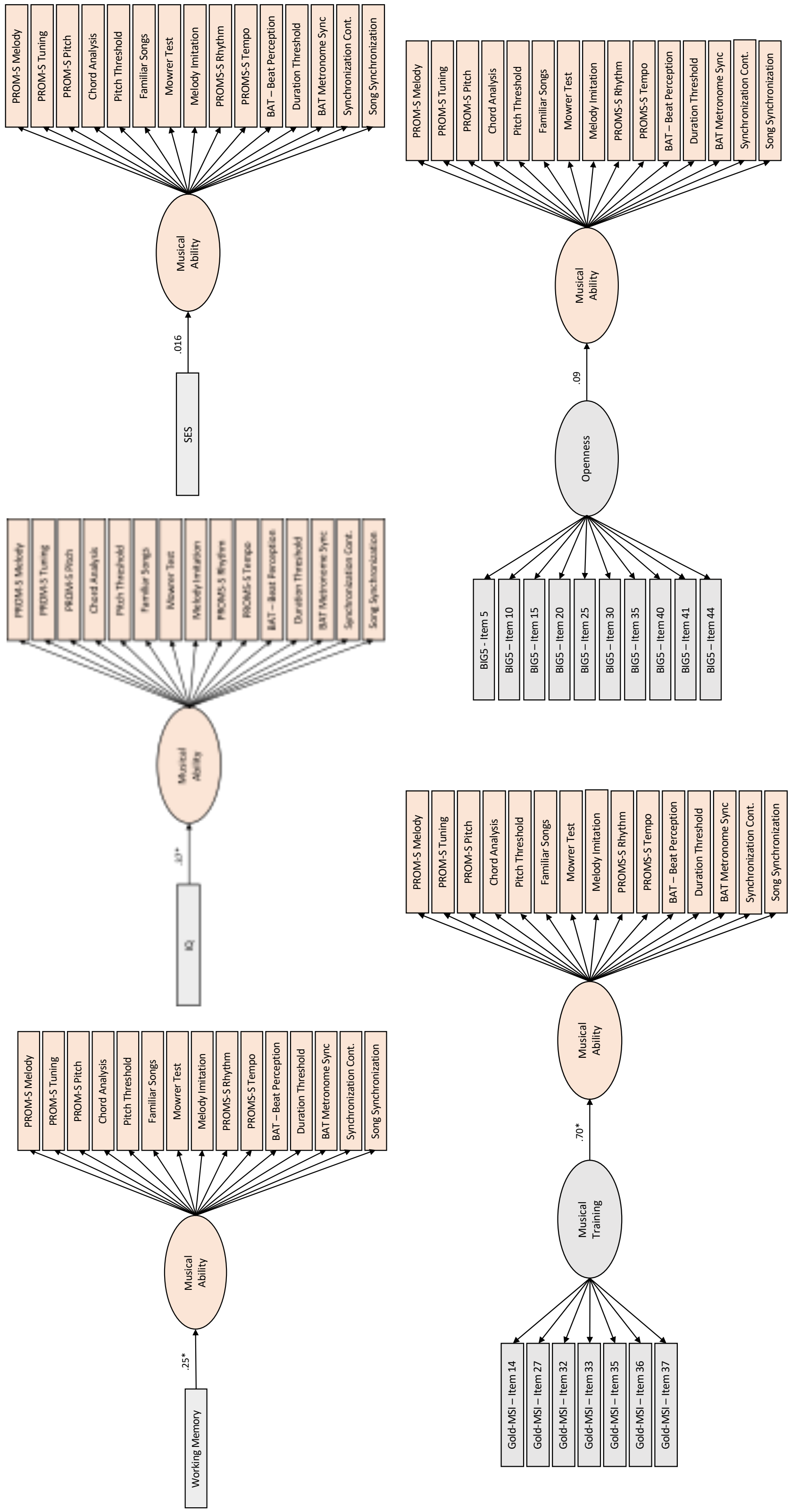


Figure 8. Individual Path Models for Ancillary Measures and Musical Ability

Chapter 4: Discussion

Although multiple tests of musical ability have been developed, there has been little consensus on what exactly constitutes musical ability and how to best measure it. Further, most musical ability tests assess only perceptual abilities with a small number of tasks. The current study used an individual differences approach and 15 representative musical ability tasks to assess whether theorized factors of musical ability (pitch, timing, perception, and production) best represented the data/performance on these tasks. Based on a series of confirmatory factor analyses, the musical ability model that best fit the data was a four-factor model including Pitch, Timing, Perception, and Production factors (Model 4A). Additionally, multiple path analyses showed that a latent measure of musical ability was positively related to musical training, working memory, and intelligence, but not to openness or socio-economic status.

The separability of Pitch and Timing factors fits with previous findings of dissociations between pitch and timing processing in amusic patients (Alcock et al., 2000; Ayotte et al., 2002; Peretz & Coltheart, 2003; Peretz et al., 2007; Phillips-Silver et al., 2011). And the relatedness of the Pitch and Timing factors fits with positive relationships found in musical ability tests measuring both pitch and timing abilities (Gordon, 2004; Peretz et al., 2003; Law & Zentner, 2012; Wallentin et al., 2010). This further supports the structure of musical ability tests in which both of these factors are assessed to estimate musical ability. The separate, but related Perception and Production factors also fit with previously found relationships where these abilities scale together (Amir & Kishon-Rabin, 2003; Dalla Bella et al., 2017; Fujii & Schlaug, 2013) and where they dissociate (Dalla Bella et al., 2011; Dalla Bella et al., 2017; Fujii & Schlaug, 2013; Loui et al., 2008). Within Model 4A, the finding that the Threshold Discrimination tasks are correlated with and are influenced by the same latent factors as more complex tasks (e.g., PROMS-S tasks) suggests that these more fine-grained auditory acuity tasks are important indicators of musical ability.

Unsurprisingly, and in relief to music educators everywhere, musical training predicted musical ability, which fits with previous studies showing this relationship (Fuji & Schlaug, 2013; Law & Zentner, 2012; Slevc et al., 2016; Swaminathan et al., 2017; Swaminathan & Schellenberg, 2018; Wallentin et al., 2010). Although the direction of this relationship is unclear, they suggest that either musical training leads to increases in musical ability and/or those who are good at musical ability tests are more likely to pursue music lessons. Within the exploratory analysis with the four-factor model, musical training predicted pitch and timing factors, but not perception nor production factors. This is interesting given that musical training increases musical exposure, so it is surprising that musical training does not predict production or perception abilities. However, it may be the case that musical training moreso influences pitch and timing abilities, which still underlie perception and production abilities on these tasks.

The positive relationship between working memory and musical ability also fits with previous findings (Okada & Slevc, 2018; Slevc et al., 2016; Swaminathan & Schellenberg, 2018; Wallentin et al., 2010) as does the positive relationship with intelligence (Swaminathan et al., 2017). This makes sense as the nature of many of the tasks used here necessitate the use of working memory in order to make same/different judgments or to listen to a musical clip then sing it aloud. However, it is interesting to note that although working memory predicted overall musical ability, it was not significantly correlated with many of the musical measures (namely the perceptual same/different judgment tasks in the PROMS-S battery). Perhaps the stimuli used for these judgments were sufficiently undemanding on memory to be within the abilities of even the low-span participants here, or participants may have recognized patterns within the stimuli (e.g., chord recognition, Povel & Jansen, 2001), so they were not as demanding on working memory processes. In regard to the exploratory analyses with the four-factor model, working memory predicted pitch and timing factors, but not perception nor production factors. Perhaps this is due to the reliance on working memory mechanisms when extracting pitch and timing aspects in relation to the overall key or meter of the musical stimulus, which may place relatively few demands on perception and production processes per se. Interestingly, the exploratory analysis with intelligence showed that intelligence predicted perception and timing factors, but not pitch nor production factors. This suggests that pattern recognition may be more important for recognizing and extracting rhythm patterns in perception tasks than in pitch or production tasks.

Surprisingly, openness to experience did not predict overall musical ability as found previously (Greenberg et al., 2015; Swaminathan & Schellenberg, 2018), both in the one factor model (see Figure 8) and four-factor exploratory analysis. This may be due to the larger battery (and larger range) of musical ability tasks assessed here, highlighting the importance of using multiple tasks to measure a construct. However, there was a significant positive correlation between openness and musical training, which does fit with previous studies on openness and musical practice and engagement (Butkovic et al., 2015; Corrigall et al., 2013). These results suggest that individuals higher in openness may be more likely to try a musical instrument and/or seek out more musical experiences, and that openness seems to be independent of musical ability.

Interestingly, SES was not related to musical ability, both in the one factor model (see Figure 8) and four-factor exploratory analysis, as has been previously found (Swaminathan et al., 2017; Swaminathan & Schellenberg, 2018). This may be due to the different measures used to assess SES here and in other studies, or may reflect the limited SES variability within our college-aged sample. However, note that previous findings show that only selective measures of SES are related to musical ability (i.e., only mother's education, but not father's education or either parental income, predicted musical ability in Swaminathan & Schellenberg, 2018), suggesting that this previously postulated relationship may not be robust. SES also did not show a relationship with musical training as seen in previous studies (Corrigall et al., 2013; Kaushal et al., 2011; Norton et al., 2005; Southgate & Roscigno, 2009). However, this non-relationship mirrors other recent findings in the same college student population (Okada & Slevc, 2018; Slevc et al., 2016), perhaps reflecting the specific sample used (see below for discussion).

Limitations and Future Directions

Despite the comprehensive set of musical tasks and the relatively large sample size used in this study, there are a few limitations. One type of limitation stems from our participant sample of University undergraduates. This specific population may have provided a restricted range of cognitive ability and SES. Of note is that our measure of SES did not correlate with any of the musical tasks nor our measure of musical training. Furthermore, it is unknown whether a different factor structure might emerge with a different population, such as with a group of highly trained musicians or those with amusia. Relatedly, it is also unknown if this factor structure would remain consistent over time or if these factors develop at different rates. Gordon (2000) asserted that musical aptitude develops until about age 9, after which it becomes stable and fixed, which suggests that this factor structure may differ for children. Hopefully future research will begin to investigate these issues.

Another type of limitation comes from the task battery. Although more comprehensive than what has been used in previous work, the tasks used in this study are just a subset of all of the existing tasks used to measure musical ability (and there is no reason to think that existing tasks fully capture all possible aspects of musical ability). In the battery used here, perhaps most notably, the pitch production tasks are a very coarse-grained view of what one could typically consider “singing.” Pitch was the only aspect measured with these data; however, future research should also examine other elements that are ascribed to good singing (e.g., appropriate vibrato, brilliance, breath management, tone quality, strain; see, e.g., Oates, Bain, David, Chapman, & Kenny, 2006), and investigate the differences between judging pitch accuracy with acoustic measurements (via software) and ratings done by music teachers (see Salvador, 2010 for a review of different rating systems). Relatedly, only one aspect of timing production was assessed here – the ability to extract the beat and tap along consistently. Future work should also consider other types of complex tapping that have been shown to vary among individuals (e.g., more complex rhythms or polyrhythms, Vuust, Wallentin, Mouridsen, Østergaard, & Roepstorff, 2011) as well as other aspects of musical ability that are likely to be related, but separate aspects from the Pitch and Timing factors seen here (e.g., timbre or loudness, as in Law & Zentner, 2012). Finally, although only basic musical ability skills were assessed here, future work will hopefully also explore the relationships between abilities like mental imagery (e.g., Halpern, Zatorre, Bouffard, & Johnson, 2004; Jakubowski et al., 2018), feeling the groove (e.g., Janata, Tomic, & Haberman, 2012), expressing and perceiving emotion in music (see Juslin & Sloboda, 2010 for a review), and other more complex aspects of engaging with music. Another limitation from the task battery (and thus also of previous studies using these tasks) is that some of the tasks have low reliabilities (i.e., Cronbach's alpha). One possible solution to this would be to remove measures that attain low reliabilities (i.e., Cronbach's alpha of less than .6 in PROMS-S Pitch, Rhythm, and Tempo subtests, Wing Chord Analysis, and BAT Perception) in estimating the best fitting model. In fact, rerunning all models without these tasks still results in the four factor model (i.e., factors of Pitch, Timing, Perception, and Production) having the best fit ($\chi^2(165) = 51.30$, p = .001, *SRMR* = .052, *RMSEA* = .076, *CFI* = .95, *AIC* = 4301.69), which suggests that this factor structure is robust. However, removing measures (or even specific items) based on reliability may undermine the established validity of the test, and additional testing should be done to ensure acceptable psychometric properties of shortened tests. Further, part of the goal of administering previously used musical tasks was to ascertain the relationships between performance on all of these tasks, and to estimate latent factors based on shared variance and minimized measurement error.

A practical limitation is that, while this study shows the importance of measuring both music perception and production abilities, analyzing production data has many notable challenges. During data collection, many participants were hesitant to sing and were embarrassed that the experimenter would hear them sing (perhaps exhibiting performance anxiety; e.g., Kenny & Osborne, 2006). A few participants sang the same melody for each trial or spoke the lyrics to the familiar songs in a monotone voice. Unfortunately, this leaves it unknown whether they possess low singing ability or if they just did not want to attempt and put forth any effort. It is also interesting that during the Melody Imitation task, some participants sang in a different octave than the musical stimuli that they heard. Although pitches from multiple octaves were judged as correct due to the perceptual similarity of pitches across octaves (i.e., octave equivalence), it may have been a harder task for those hearing notes in a given octave, then producing notes in a different octave. Although work on pitch recall for pitches heard in different octaves shows that accuracy remains similar despite octave displacement (Deutsch & Boulanger, 1984), it is unknown whether they were aware that they were singing in a different octave from what they heard and if doing so was more difficult. Furthermore, singing recordings can be a very noisy signal, and programs designed to extract pitch frequency may not always do so accurately (e.g., pitch tracking errors may occur if or if optimal parameters are not set, Babacan, Drugman, d'Alessandro, Henrich, & Dutoit, 2013; Murray, 2001). Future work should try to use robust pitch extractors to minimize these types of errors.

Advancing this model of musical ability that takes into account relevant factors (pitch, timing, perception and production) will hopefully be an important contribution to the field of music psychology. To date, most studies only consider perceptual abilities even though the main criterion outcome of interest from these tests is whether or not someone will succeed in learning/producing music. The separability between these four factors is evidence that musical ability should be measured with a range of tasks. Future research on musical ability should attempt to measure production aspects as well as perception aspects and also be more explicit about what performance on tasks actually represents (i.e., performance on one perceptual task should not generalize to someone's overall musical ability).

As a methodological contribution, another goal of this project was to encourage the use of more sophisticated individual differences approaches to investigate musical ability and its relationships to other constructs. Using multiple tasks to measure a given construct as well as using latent variable analysis provided more robust results to compare to the existing relationships found in the literature. With clearer reporting of measures used, the field will hopefully move away from mixed findings and toward a clearer, more nuanced picture of musical ability and its relationships with other measures.

Finally, another future goal of this project is to develop a new, shorter test battery (with multiple tasks measuring each factor) that will use freely available materials that can reliably measure pitch and timing perception and production abilities in the general population. This newer battery could be used by music researchers investigating musical ability in adults or by music educators to ascertain ability levels of their students. Within the best fitting model here, the four latent factors predicted the most variance in the Mowrer Test of Tonal Memory ($R^2 = .84$), Metronome Synchronization task ($R^2 = .69$), Pitch Threshold task ($R^2 = .64$), BAT Perception task ($R^2 = .45$) and Duration Threshold task ($R^2 = .44$). Interestingly, the Mowrer Test of Tonal Memory has the highest R^2 value, which underscores its use to predict choral ensemble contribution (Mowrer, 1996), and is a quick test that can easily be administered by music educators or researchers to measure this aspect of music ability. The fact that the tasks with the highest R^2 values also highlight that pitch and timing production as well as finer-grained perception are critical in measuring musical ability.

In sum, although musical ability has been extensively studied, most conclusions about the nature of musical ability and its relationships with other factors have been drawn using limited sets of mainly perceptual musical ability tasks. By using a more comprehensive set of musical ability tasks and evaluating structural latent variable models, this work will hopefully advance research investigating different aspects of musical ability and provide a stronger framework for future studies investigating the complex nature of musical ability.

Appendices

Appendix A

Disattenuated Correlation Matrix of Musical Measures

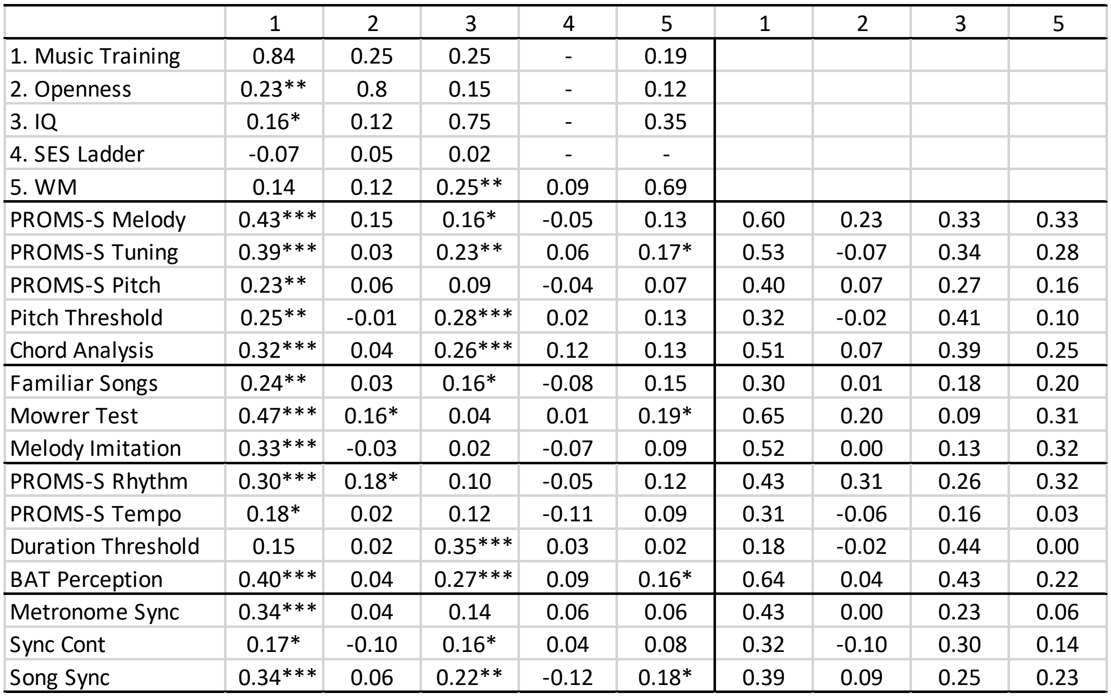
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pitch Perception	1. PROMS-S Melody	0.61	0.53	0.54	0.39	0.43	0.24	0.74	0.42	0.47	0.25	0.21	0.58	0.57	0.21	0.26
	2. PROMS-S Tuning	0.32***	0.65	0.57	0.38	0.56	-0.04	0.57	0.29	0.48	0.45	0.27	0.41	0.31	0.21	0.37
	3. PROMS-S Pitch	0.27***	0.27***	0.46	0.36	0.27	0.05	0.61	0.38	0.36	0.29	0.41	0.51	0.41	0.21	0.26
	4. Pitch Threshold	0.23**	0.23**	0.22**	0.85	0.54	0.14	0.34	0.13	0.27	0.2	0.64	0.37	0.57	0.43	0.21
	5. Chord Analysis	0.23**	0.28***	0.11	0.35***	0.53	0.12	0.57	0.39	0.34	0.07	0.4	0.49	0.45	0.37	0.39
Pitch Production	6. Familiar Songs	0.17*	-0.03	0.08	0.18*	0.11	0.64	0.28	0.51	0.16	-0.07	-0.06	0.12	0.22	0.23	-0.05
	7. Mowrer Test	0.51***	0.36***	0.37***	0.24**	0.33***	0.19*	0.77	0.64	0.47	0.38	0.1	0.64	0.47	0.31	0.41
	8. Melody Imitation	0.30***	0.13	0.20*	0.06	0.24**	0.32***	0.53***	0.8	0.12	0.17	0.04	0.37	0.15	0.13	0.16
Timing Perception	9. PROMS-S Rhythm	0.29***	0.26***	0.19*	0.16*	0.14	0.13	0.28***	0.06	0.52	0.38	0.31	0.4	0.39	0.26	0.41
	10. PROMS-S Tempo	0.16*	0.24**	0.13	0.17*	0.05	-0.03	0.26***	0.07	0.20*	0.49	0.37	0.5	0.41	0.2	0.42
	11. Duration Threshold	0.20*	0.23**	0.19*	0.46***	0.26**	-0.02	0.14	0.03	0.17*	0.26**	0.75	0.55	0.44	0.35	0.25
	12. BAT Perception	0.29***	0.24**	0.26***	0.28***	0.27***	0.06	0.41***	0.17*	0.18*	0.31***	0.39***	0.56	0.54	0.43	0.45
Timing Production	13. Metronome Sync	0.34***	0.21**	0.20*	0.49***	0.31***	0.17*	0.31***	0.09	0.21**	0.27***	0.33***	0.36***	0.75	0.7	0.39
	14. Sync Cont	0.10	0.09	0.11	0.34***	0.23**	0.14	0.20**	0.11	0.12	0.17*	0.22**	0.29***	0.50***	0.62	0.46
	15. Song Sync	0.19*	0.27***	0.14	0.16*	0.25**	-0.03	0.28***	0.07	0.25**	0.27***	0.20*	0.33***	0.28***	0.29***	0.93

Note. **p* < .05, ***p* < .01, ****p* < .001

Pearson's correlations are below the diagonal, reliabilities for each task are on the diagonal, and disattenuated correlations are above the diagonal. Cronbach's alpha was used here because it was not possible to estimate Omega for all measures.

Appendix B

Disattenuated Correlation Matrix of Ancillary Measures and Musical Measures



	1	2	3	4	5	1	2	3	5
1. Music Training	0.84	0.25	0.25	-	0.19				
2. Openness	0.23**	0.8	0.15	-	0.12				
3. IQ	0.16*	0.12	0.75	-	0.35				
4. SES Ladder	-0.07	0.05	0.02	-	-				
5. WM	0.14	0.12	0.25**	0.09	0.69				
PROMS-S Melody	0.43***	0.15	0.16*	-0.05	0.13	0.60	0.23	0.33	0.33
PROMS-S Tuning	0.39***	0.03	0.23**	0.06	0.17*	0.53	-0.07	0.34	0.28
PROMS-S Pitch	0.23**	0.06	0.09	-0.04	0.07	0.40	0.07	0.27	0.16
Pitch Threshold	0.25**	-0.01	0.28***	0.02	0.13	0.32	-0.02	0.41	0.10
Chord Analysis	0.32***	0.04	0.26***	0.12	0.13	0.51	0.07	0.39	0.25
Familiar Songs	0.24**	0.03	0.16*	-0.08	0.15	0.30	0.01	0.18	0.20
Mowrer Test	0.47***	0.16*	0.04	0.01	0.19*	0.65	0.20	0.09	0.31
Melody Imitation	0.33***	-0.03	0.02	-0.07	0.09	0.52	0.00	0.13	0.32
PROMS-S Rhythm	0.30***	0.18*	0.10	-0.05	0.12	0.43	0.31	0.26	0.32
PROMS-S Tempo	0.18*	0.02	0.12	-0.11	0.09	0.31	-0.06	0.16	0.03
Duration Threshold	0.15	0.02	0.35***	0.03	0.02	0.18	-0.02	0.44	0.00
BAT Perception	0.40***	0.04	0.27***	0.09	0.16*	0.64	0.04	0.43	0.22
Metronome Sync	0.34***	0.04	0.14	0.06	0.06	0.43	0.00	0.23	0.06
Sync Cont	0.17*	-0.10	0.16*	0.04	0.08	0.32	-0.10	0.30	0.14
Song Sync	0.34***	0.06	0.22**	-0.12	0.18*	0.39	0.09	0.25	0.23

Note. **p* < .05, ***p* < .01, ****p* < .001
In the top left quadrant with the five ancillary measures, Pearson's correlations are below the diagonal, reliabilities for each task are on the diagonal, and disattenuated correlations are above the diagonal.
For correlations with the musical tasks, Pearson's correlations are on the left and disattenuated correlations are on the right. Cronbach's alpha was used here because it was not possible to estimate Omega for all measures. There are no disattenuated correlations with SES because reliability could not be calculated for this measure.

References

Adler, N., & Stewart, J. (2007). The MacArthur scale of subjective social status. *MacArthur Research Network on SES & Health. Retrieved from http://www.macses.ucsf.edu/Research/Psychosocial/subjective.php.*

Albin, A. (2014). PraatR: An architecture for controlling the phonetics software “Praat” with the R programming language. *Journal of the Acoustical Society of America, 135*(4), 2198.

Alcock, K. J., Passingham, R. E., Watkins, K., & Vargha-Khadem, F. (2000). Pitch and timing abilities in inherited speech and language impairment. *Brain and language*, *75*(1), 34-46.

Amir, O., Amir, N., & Kishon-Rabin, L. (2003). The effect of superior auditory skills on vocal accuracy. *Journal of the Acoustical Society of America, 113*(2), 1102-1108.

Audacity Team (2010). Audacity®: Free Audio Editor and Recorder [Computer program]. Version 1.3.12 retrieved from https://audacityteam.org.

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain*, *125*(2), 238-251.

Babacan, O., Drugman, T., d’Alessandro, N., Henrich, N., & Dutoit, T. (2013, May). A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 7815-7819).

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

Benz, S., Sellaro, R., Hommel, B., & Colzato, L. S. (2015). Music makes the world go round: The impact of musical training on non-musical cognitive functions - A review. *Frontiers in Psychology, 6*, 1-5.

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience*, *23*(2), 425-434.

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer [Computer program]. Version 6.0.19 retrieved from http://www.praat.org.

Bugos, J. A., & Kochar, S. (2017). Efficacy of a short-term intense piano training program for cognitive aging: A pilot study. *Musicae Scientiae, 21*(2), 137-150.

Bugos, J. A., Perlstein, W. M., McCrae, C. S., Brophy, T. S., & Bedenbaugh, P. H. (2007). Individualized piano instruction enhances executive functioning and working memory in older adults. *Aging & Mental Health, 11*(4), 464–71.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge: New York.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Moving on time: Brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of Cognitive Neuroscience, 20*(2), 226-239.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Conard, N. J., Malina, M., & Münzel, C. (2009). New flutes document the earliest musical tradition in southwestern Germany. *Nature: Letters*, *460*(6), 737-740.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163-183.

Corrigall, K. A., & Schellenberg, E. G. (2015). Predicting who takes music lessons: Parent and child characteristics. *Frontiers in Psychology*, *6*(282), 1-8.

Corrigall, K. A., Schellenberg, E. G., & Misura, N. M. (2013). Music training, cognition, and personality. *Frontiers in Psychology*, *4*(222), 1-10..

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Dalla Bella, S., Berkowska, M., & Sowiński, J. (2011). Disorders of pitch production in tone deafness. *Frontiers in Psychology*, *2*(146), 1-11.

Dalla Bella, S., Farrugia, N., Benoit, C. E., Begel, V., Verga, L., Harding, E., & Kotz, S. A. (2017). BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities. *Behavior Research Methods*, *49*(3), 1128-1145.

Dediu, D., & Levinson, S. C. (2013). On the antiquity of language: The reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in Psychology, 4*(397), 1-17.

Deutsch, D. (2013). The Processing of Pitch Combinations. In D. Deutsch (Ed.), *Psychology of Music* (3rd ed.) (pp. 268-345). San Diego, CA: Academic Press.

Deutsch, D., & Boulanger, R. (1984). Octave equivalence and the immediate recall of pitch sequences. *Music Perception, 2*(1), 40-51.

Elpus, K. (2013). Is it the music or is it selection bias? A nationwide analysis of music and non-music students' SAT scores. *Journal of Research in Music Education, 61*, 175-194.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309.

Ericsson, K. A., Krampe, R. T., Tesch-Romer, C., (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Fujii, S., & Schlaug, G. (2013). The Harvard Beat Assessment Test (H-BAT): A battery for assessing beat perception and production and their dissociation. *Frontiers in Human Neuroscience, 7,* 1-16.

Gordon, E. E. (1965). The musical aptitude profile: A new and unique musical aptitude test battery. *Bulletin of the Council for Research in Music Education*, 12–16.

Gordon, E. E. (1989). *Advanced Measures of Music Audiation.* Chicago: Riverside Publishing Company.

Gordon, E. E. (1995). Manual Musical Aptitude Profile (Third Edition.). Chicago: GIA Publications, Inc.

Gordon, E. E. (2001). Music Aptitude and Related Tests- An Introduction. Chicago: GIA Publications, Inc.

Gordon, E. E. (2004). Continuing Studies in Music Aptitudes. Chicago: GIA Publications, Inc.

Grassi, M., & Soranzo, A. (2009). MLP: A Matlab toolbox for rapid and reliable auditory threshold estimation. *Behavior Research Methods, 41*(1), 20-28.

Habibi, A., Damasio, A., Ilari, B., Sachs, M. E., & Damasio, H. (2018). Music training and child development: A review of recent findings from a longitudinal study. *Annals of the New York Academy of Sciences, 1423*, 73-81.

Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Sciences*, *13*(2), 65-73.

Hallam, S., & Prince, V. (2003). Conceptions of Musical Ability. *Research Studies in Music Education, 20*(1), 2-22.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia 42*, 1281-1292.

Hermann, E. (1981). Shinichi Suzuki: The Man and His Philosophy. Miami, FL: Summy-Birchard Inc.

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.

Iversen, J. R, & Patel, A. D. (2008). The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In K. Miyazaki, M. Adachi, Y Hiraga, Y Nakajima, & M. Tsuzaki (Eds.), *Proceedings of the 10th*

International Conference on Music Perception & Cognition (ICMPC10) (CD-ROM; pp. 465–468). Adelaide, Australia: Causal Productions.

Janata, P., Tomic, S. T., & Haberman, J. M. (2012). Sensorimotor coupling in music and the psychology of the groove. *Journal of Experimental Psychology: General, 141*(1), 54-75.

John, O. P., Donahue, E. M., Kentle, R. L. (1991). The Big Five Inventory – Versions 4a and 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (pp. 114-158). New York, NY: Guilford Press.

Juslin, P. N., & Sloboda, J. A., (Eds.) (2010). *Handbook of music and emotion: Theory, research, applications.* New York: Oxford University Press.

Kenny, D. T., & Osborne, M. S. (2006). Music performance anxiety: New insights from young musicians. *Advances in Cognitive Psychology, 2*(2-3), 103-112.

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418-435.

Koelsch, S. (2011). Toward a neural basis of music perception - a review and updated model. *Frontiers in Psychology, 2*, 1-20.

Koelsch, S., Gunter, T. C., von Cramon, D. Y., Zysset, S., Lohmann, G., & Friederici, A. D. (2002). Bach speaks: A cortical" language-network" serves the processing of music. *Neuroimage*, *17*(2), 956-966.

Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Science, 9*(12), 578-584.

Koza, J. E., (2007). In sounds and silences: Acknowledging political engagement. *Philosophy of Music Education Review, 15*(2), 168-176.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599-605.

Law, L. N. (2012). *Assessing and Understanding Individual Differences in Music Perception Abilities* (Doctoral dissertation). Retrieved from http://etheses.whiterose.ac.uk/3142/2/THESIS_LilyLaw_FINAL.pdf

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS one*, *7*(12), 1-15.

Loui, P., Guenther, F. H., Mathys, C., & Schlaug, G. (2008). Action–perception mismatch in tone-deafness. *Current Biology*, *18*(8), 331-332.

McAuley, J. D. (2010). Tempo and rhythm. In M. R. Jones, R. R. Fay, & A. R. Popper (Eds), *Music Perception: Springer Handbook of Auditory Research* (pp. 165-200), New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Hillsdale: Lawrence Erlbaum Associate.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cognitive Psychology, 41*(1), 49-100.

Moore, R. E., Estis, J., Gordon-Hickey, S., & Watts, C. (2008). Pitch discrimination and pitch matching abilities with vocal and nonvocal stimuli. *Journal of Voice*, *22*(4), 399-407.

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

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

Mowrer, A. (1996). *Tonal Memory as an Audition Factor for Choral Ensembles* (Doctoral dissertation). Retrieved from personal communication.

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), 1-23.

Murdoch, S. (2007). IQ: A smart history of a failed idea. Hoboken, NJ: John Wiley.

Murray, K. (2001, September). A study of automatic pitch tracker doubling/halving errors. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue – Volume 16* (pp. 1-4). Association for Computational Linguistics.

Norris, C. E., (2000). Factors related to the validity of reproduction tonal memory tests. *Journal of Research in Music Education, 48*(1), 52-64.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7(*6), 615-631.

Oates, J. M., Bain, B., Davis, P., Chapman, J., & Kenny, D. (2006). Development of an auditory-perceptual rating instrument for the operatic singing voice. *Journal of Voice, 20*(1), 71-81.

Okada, B. M., & Slevc, L. R. (2018). Individual differences in musical training and executive funtions: A latent variable approach. *Memory & Cognition,* 1-11.

Okada, B. M., & Slevc, L. R. (in press). Musical training: Contributions to executive function. In M. Bunting, J. Novick, M. Dougherty, & R. W. Engle (Eds.), *An integrative approach to cognitive and working memory training: Perspectives from psychology, neuroscience, and human development.* New York, NY: Oxford University Press. https://doi.org/10.13016/M2GM81P70

Oxenham, A. J. (2013). The perception of musical tones. In D. Deutsch (Ed.), *Psychology of Music* (3rd ed.) (pp. 20-53). San Diego, CA: Academic Press.

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8-13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*(10), 1-8.

Peretz, I. (2001). Brain specialization for music: New evidence from congenital amusia. *Annals of the New York Academy of Sciences*, *930*(1), 153-165.

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.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, *6*(7), 688-691.

Peretz, I., Cummings, S., & Dubé, M. P. (2007). The genetics of congenital amusia (tone deafness): A family-aggregation study. *The American Journal of Human Genetics*, *81*(3), 582-588.

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception: An Interdisciplinary Journal*, *25*(2), 95-115.

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M. Liotti, M. (2010). Imprecise singing is widespread. *Journal of the Acoustical Society of America, 128*(4), 2182-2190.

Phillips-Silver, J., Toiviainen, P., Gosselin, N., Piché, O., Nozaradan, S., Palmer, C., & Peretz, I. (2011). Born to dance but beat deaf: A new form of congenital amusia. *Neuropsychologia*, *49*(5), 961-969.

Povel, D., & Jansen, E., (2001). Perceptual mechanism in music processing. *Music Perception, An Interdisciplinary Journal, 19*(2), 169-197.

R Development Core Team. (2016). *R: A language and environment for statistical computing* [Computer program]. Vienna, Austria: R Foundation for Statistical Computing.

Raven, J. C. , Raven, J., & Court, J. H. (1991). Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 1). Oxford, UK: Oxford Psychologists Press.

Revelle, W. (2018). Using R and the psych package to find ω [PDF file]. Retrieved from http://personality-project.org/r/psych/HowTo/omega.pdf

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145-154.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Salvador, K. (2010). How can elementary teachers measure singing voice achievement? A critical review of assessments, 1994-2009. *Update: Applications of Research in Music Education, 29*(1), 40-47.

Sattora, A., & Bentler, P. M. (2001). A scaled chi-square difference test for moment structure analysis. *Psychometrika, 66*, 507-514.

Schellenberg, E. G. (2004). Music Lessons Enhance IQ. *Psychological Science, 15*(8), 511–514.

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In D. Deutsch (Ed.), *Psychology of Music* (3rd ed.) (pp. 499-550). San Diego, CA: Academic Press.

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

Seashore, C. E. (1919). The psychology of musical talent. Boston: Silver, Burdett, and Company.

Seashore, C. E. (1938). Psychology of Music. New York: McGraw-Hill Book Company.

Seashore, C. E., Lewis, D., & Saetveit, J. G. (1960). Seashore Measures of Musical Talent's Manual (Revised). New York: Psychological Corporation.

Shipstead, Z., Lindsey, D. R., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language, 72*, 116-141.

Shuter, R. (1966). Hereditary and environmental factors in musical ability. *The Eugenics Review, 58*(3), 149-156.

Singh-Manoux, A., Marmot, M. G., & Adler, N. E. (2005). Does subjective social status predict health and change in health status better than objective status? *Psychosomatic Medicine, 67*, 855-861.

Slevc, L.R., Davey, N., Buschkuehl, M., & Jaeggi, S.M. (2016). Tuning the mind: Exploring the connections between musical ability and executive functions. *Cognition, 152*, 199-211.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science*, *17*(8), 675-681.

Soranzo, A., & Grassi, M. Psychoacoustics: A comprehensive Matlab toolbox for auditory testing. *Frontiers in Psychology, 5*(712), 1-13.

Southgate, D. E., & Roscigno, V. J. (2009). The impact of music on childhood and adolescent achievement*. *Social Science Quarterly, 90*(1), 4-21.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101.

Stanton, H. (1922). Inheritance of specific musical capacities. *Psychological Monographs, 31,* 157-204.

Suzuki, S. (1981). *Ability development from age zero.* (M. L. Nagata, Trans.) Miami, FL: Warner Brothers Publications, Inc.

Swaminathan, S., & Schellenberg, E. G. (2018). Musical competence is predicted by music training, cognitive abilities, and personality. *Nature Scientific Reports, 8*(1), 1-7.

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.

Tan, Y. T., McPherson, G. E., Peretz, I., Berkovic, S. F., & Wilson, S. J. (2014). The genetic basis of music ability. *Frontiers in Psychology*, *5*(658), 1-19.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language, 69*, 36-58.

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.

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale III*. San Antonio, TX: Harcourt Assessment.

Whiteford, K., & Oxenham, A. J. Learning for pitch and melody discrimination in congenital amusia. *Cortex, 103*, 164-178.

Wing, H. D. (1948). Tests of musical ability and appreciation. An investigation into the measurement, distribution, and development of musical capacity. *The British Journal of Psychology*. Monograph Supplements. London: Cambridge University Press.

Wing, H. D. (1962). A Revision of the "Wing Musical Aptitude Test". *Journal of Research in Music Education, 10*(1), 39-46.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*(1), 165-200.

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences, 1400,* 33-45.