Introduction

Once largely confined to the domain of music education, today’s interest in musical ability is increasingly driven by research questions in neuroscience, medicine, and psychology. One reason for this attraction to research of the study of musical ability is that it is associated with a number of nonmusical abilities, including reading, phonological awareness, second language abilities, mathematical ability, spatial abilities, memory, executive functions, and motor skills (see Schellenberg, 2016a, for a review). These associations are studied not only to advance our understanding of mental and neural structures, but also to explore the potential of enhancing musical ability as a form of therapeutical treatment. There is now increasing evidence that music training may help to restore normal functioning in people suffering from psychological or medical conditions, such as dementia, dyslexia, or autism spectrum disorder (e.g., Chanda & Levitin, 2013, Chapter 14 for reviews).

The credibility of this growing body of findings is contingent on the methods used for assessing musical ability. At present, researchers generally adopt one of two approaches. The first, and most widely used, consists of classifying participants as musicians or nonmusicians and comparing the two groups on some outcome of interest (Schellenberg, 2016a). When musicians outperform nonmusicians, the result is interpreted as evidence for a link between musical ability and the outcome being studied, such as working memory or spatial ability (e.g., Hansen, Wallentin, & Vuust, 2013; Sluming, Brooks, Downes, & Roberts, 2007). Because professional musicians are in general more musically capable than nonmusicians, this approach is reasonable. Nevertheless, the interpretation of findings obtained with it is not straightforward.

First, being a “nonmusician” does not, in and of itself, denote an absence of musical ability. The ability may be undiscovered, or circumstances may have prevented its development. To account for these individuals with latent musical ability, a binary classification based
on musicianship status is clearly not sufficient. A second difficulty is that specific musical skills are difficult to assess by means of comparing musicians and nonmusicians. One might compare groups of different musicians that conceivably require different musical skills, such as drummers and singers (Slater & Kraus, 2016), but such a procedure infers specific musical skills indirectly rather than providing a direct assessment thereof.

A second approach to the measurement of musical ability is by means of standardized musicality tests. Tests have some advantages over the classification approach described above. For example, because some musically untrained individuals are presumably more musically gifted than others, tests can help to identify individual differences in musical skills among the musically untrained. Moreover, specific skills such as rhythmic or tonal abilities can be assessed with most existing musical aptitude tests, thus allowing for more fine-grained analyses and interpretations. Standardized tests can also help determine whether benefits of musical training are due to improvements in musical abilities or to other features of musical training. At the same time, the usefulness of tests heavily depends on the extent to which they have been validated. For instance, if a test of musical ability were to measure working memory as well, any link between working memory and musical ability test scores would be trivial. For this reason, the psychometric characteristics of musical aptitude tests warrant particular scrutiny and will therefore be given serious consideration in this chapter.

The aim of this chapter is twofold. The first is to provide a working definition of musical ability and review test batteries for the assessment of musical ability. Because these batteries were often conceived as instruments for identifying aptitude (i.e., musical potential that is inherited or at least present at birth), a second goal is to review the evidence relating to the heritability and genetics of musical ability. Finally, we provide a critical appraisal of current views and measures of musical ability and offer some suggestions for future research.

**The Concept of Musical Ability**

Various terms are used in the specialized literature to refer to abilities in the domain of music, including *musical giftedness, musical talent, musical aptitude, musical ability, musical capacity, musical intelligence, musical achievement, musical accomplishment, and musicality* (Radocy & Boyle, 2012). Running through the plethora of terms is a basic distinction between potential and actual levels of musical ability, usually referred to as a difference between *musical aptitude* and *musical achievement*.

The term *musical aptitude* has been defined as the potential for learning music that is present before formal training and achievement (Shuter-Dyson, 1999). Consider two individuals with similar background, education, intelligence, and personality dispositions who are given the same musical training. Should the two students differ in their ease or speed of acquiring
musical skills, such differences would be indicative of differences in musical aptitude. In contrast, musical attainment (or achievement) refers to the actual level of musical accomplishment as shown at any given point in life, regardless of its etiology. In the sense that reading, performing, or composing music would be virtually impossible without musical instruction and training, musical attainment involves acquired skills. Musical ability is the more general term encompassing aptitude as well as attainment (see also Radocy & Boyle, 2012).

The main implication of these terminological distinctions for assessment purposes is that musical aptitude can be assessed irrespective of a person’s previous level of musical instruction. If musicality is understood in terms of aptitude, then nonmusicians can be more or less musical. Nonmusicians with substantial musical aptitude may be able to reach high levels of musical ability, given the time and opportunity to develop their talents at a young age. Because their skills are dormant, we refer to these individuals as musical sleepers. Sleeping musicians, in turn, are individuals whose musical proficiency languishes despite multiple years of training, degrees, and certificates (Law & Zentner, 2012). Although contrasting musical sleepers with sleeping musicians fails to capture many intermediate possibilities, the metaphor is helpful in pinpointing the main shortcoming of using musicianship as a proxy for musical ability.

With these preliminaries regarding meaning of ability clarified, we can turn to questions about the contents and dimensions of musical ability. Because attempts to define the elements of musicality go back to the nineteenth century (e.g., Billroth, 1895; Michaelis, 1805), providing an extensive coverage of different views on the subject would be beyond the scope of this chapter. Let it just be said here that most scholars see musical ability as a capacity that includes several components. For example, Shuter-Dyson holds that at least five groups of fundamental musical abilities can be distinguished: tonal, rhythmic, kinesthetic, aesthetic, and creative, each of which could be subdivided into other subcomponents, all of which are subject to improvement with age and exposure (Shuter-Dyson & Gabriel, 1981; Shuter-Dyson, 1999).

More recently, Levitin (2012) has argued in favor of a more detailed classification of musical behaviors and aptitudes, one that might also serve to describe differences in established professionals: (a) playing a musical instrument or singing; (b) composing; (c) arranging and orchestrating; (d) conducting; (e) programming music for aesthetic purposes or for finding connections between songs (disc jockeys, film supervisors); (f) receptive sensitivity to music and its emotional content; (g) ability to detect out-of-tune or out-of-key notes; (h) crossmodal practices, such as writing about or choreographing music.

For our purposes, it is sufficient to highlight some broad distinctions that have particular relevance in the context of assessment practices. Here, we will classify tools for the assessment of musicality in three broad categories: instruments for the assessment of (a) abilities
in music perception, (b) abilities in music production, and (c) the capacity for enjoyment of music. Abilities in *music perception* relate to the ability to hear, such as the capacity to identify a note that is out of tune, to discriminate between two instrumental sounds, or to perceive subtle differences in the expressiveness of different performances of the same music. Abilities in *music production* relate to abilities in creating and performing music, such as composing, arranging, improvising, or performing.

The capacity for the *enjoyment of music* relates to individual differences in the ability to derive pleasure from music. Whereas some people can live happily without music, for others, a life without music would feel greatly impoverished. Although the capacity for musical enjoyment is not an ability in a strict sense, we include it in our classification as an area that deserves more attention than it has so far received. For example, is the enjoyment of music related to abilities in the perception or production of music? Is it associated with wellbeing or particular personality attributes?

In the context of assessment tools and practices, distinguishing between perceptual-receptive skills on one hand (i.e., capacities in the perceptual and the hedonic sphere) and production skills (i.e., playing a musical instrument) on the other has particular significance because the former can be measured virtually in the entire population, regardless of people’s musical background or training. In contrast, most production abilities require musical training, thereby confining the target population to a minor portion of the population. According to a survey conducted in Germany, professional musicians make up less than 1 percent of the adult population (Klein, 2011).

At present, production skills are most commonly evaluated by experts in the context of music exams or competitions, using criteria that often vary from situation to situation. Exams and competitions also require preparation, whereas the opposite is the case for aptitude tests, as noted by Sloboda (1985): “Whilst examinations presuppose intensive preparation of specific materials, tests of ability involve no foreknowledge of test content. Indeed, such tests are invalidated by extensive practice on the task they contain” (p. 233).

Although most musical production skills require formal musical training, there are some exceptions. For example, tapping and singing are produced spontaneously by young children and do not seem to require formal musical training (e.g., Trehub & Gudmundsdottir, 2015). Thus, tasks measuring individual differences in these musical behaviors could perhaps be included in batteries for the standardized assessment of musical ability. Although some measures of tapping and singing do already exist (e.g., Berkowska & Dalla Bella, 2013; Hutchins & Peretz, 2012; Iversen & Patel, 2008), most of them were conceived ad hoc, in the context of a specific research project, rather than as tests for the ubiquitous assessment of musical ability (but see Dalla Bella et al., 2016 for an exception). Another potential challenge of production skill measures lies in their technical demands for administration and scoring. For example,
measures of tapping skills require precise reaction times to be recorded. In the context of studies that are conducted in the laboratory, this is feasible, but it may be more difficult to ensure if the tasks are to be delivered online.

Online administration of musical ability measures offers significant practical advantages, especially if researchers want to study large or special populations. The population range to which musical assessments may be applied has enormous significance, because the inclusion of large strata of the population enables researchers to (a) recruit large numbers of participants and thus achieve higher statistical power; (b) obtain findings that can be generalized more safely; (c) study musical abilities in rare or special subpopulations, such as autistic children or patients with dementia; and (d) conduct cross-cultural studies without the challenges involved in recruiting large numbers of practicing musicians.

In summary, musical ability is a construct with many facets, some of which are more amenable to standardized testing than others. Despite some promising initiatives (e.g., Dalla Bella et al., 2016), overall, tests of musical production capacities that could be used with untrained populations are not advanced enough to warrant detailed coverage in the following review of tests. Similarly, individual differences in receptivity to music-induced pleasure have not been widely studied (but see Mas-Herrero, Marco-Pallares, Lorenzo-Seva, Zatorre, & Rodriguez-Fornells, 2013, for an exception). For the time being, such differences may be assessed by means of music appreciation questionnaires. The situation is different for measures of music perception skills, which are comparatively well developed and will therefore be the focus of the next section.

**Review of Musical Ability Tests**

*Introductory Comments about Psychological Testing*

To readers unacquainted with principles of psychological testing and measurement, we provide a few introductory remarks so as to facilitate the understanding of the subsequent sections. The knowledgeable reader may skip this section.

Psychological tests typically involve a series of tasks or problems that the respondent has to solve. They can be used in the context of individual testing, as part of selection processes (e.g., job interviews, to assess children in schools, to assess people with mental health issues, or to assess offenders in prisons). In research contexts, they are more typically used as a means of quantifying people’s traits or skills so that the skills and traits can be studied in relation to specific precursors or outcomes or to other mental and/or neurobiological attributes. For example, working memory tests may be administered to hundreds of school children to understand how working memory relates to other mental characteristics, such as linguistic abilities. They may also be administered to examine how working memory capacity relates...
to predictors such as genetic and environmental factors or to outcomes such as educational or occupational attainment.

In what follows we will discuss musical ability tests in the latter sense—that is, as research instruments rather than as diagnostic tools for individual assessment and selection processes. To serve their role as research instruments, tests must meet a number of quality criteria. To discuss these criteria in any detail would be beyond the scope of this chapter. A few remarks about the most frequently mentioned ones, validity and reliability, will have to suffice.

Validity relates to different forms of correspondence between a test and the phenomenon it intends to measure. One form of correspondence, criterion validity, is the extent to which test performance and performance outside the testing context are meaningfully related. For example, a test of reading comprehension may be said to be valid if performance on the test corresponds to reading grades obtained in school. Another form, convergent validity, is the correspondence between the performance on a given test and the performance on similar tests with an already established record of validity. Discriminant validity refers to the specificity of correspondence, that is, the extent to which a test measures the specific ability it purports to measure rather than some other ability. A test of verbal comprehension that predicts language proficiency as much as it predicts short-term memory would have poor discriminant validity. A test may also be inspected with regard to content validity to ascertain whether the content of the test is appropriate for its intended purpose. For example, content validity would be poor if a test of mathematical ability included only arithmetic tasks.

Reliability refers to consistency of measurement. It is often evaluated by two indicators, internal consistency and test-retest reliability. Internal consistency reliability reflects the degree to which test items that propose to measure the same ability produce similar scores. The degree of similarity is often indexed through a coefficient called Cronbach alpha (α). Test-retest reliability is present when a test produces consistent results across separate testing sessions (for more information on principles of psychological testing the reader is referred to Kline, 2000).

The Evolution of Musical Ability Tests
Although ideas about the definition and evaluation of musicality go back to the nineteenth century, and tasks for the assessment of musicality were developed by Stumpf (1909) and Revesz (1920), standardized, psychometrically controlled tests of musical ability were pioneered by Carl E. Seashore. As early as 1912, Seashore presented tasks for the assessment musical ability before a women’s club in Des Moines, Iowa (Miles, 1956), but the complete test battery was not published until several years later (Seashore, 1919). Numerous musical ability tests were published over the next half-century, including those by Kwalwasser & Dykema.
(1930), Tilson (1941), Drake (1954), Gaston (1957), Bentley (1966), Wing (1968), Gordon (1965, 1989), and Karma (1973, 1975; see also Shutet-Dyson & Gabriel, 1981, and Treichler, 2013, for a review of these batteries.)

Remarkably, work on musical ability testing came virtually to a halt after the 1970s. Except for developments of batteries that had already been created in 1960s and 1970s and a few initiatives that remained largely undeveloped (Mills, 1988; Cohen, 1992), hardly any new musical aptitude tests were published between 1980 and 2000. It is not until the early 2000s that work on standardized tests of musical ability resumed. In recognition of this rift, we shall refer to the test developed before the gap as the earlier tests and to those published after 2000 as the newer ones.

More than anything else, it is the aim that distinguishes the earlier tests from the newer ones. Most of the earlier tests were conceived in the context of music education, as tools to identify children deemed sufficiently gifted to receive a formal music education. In the revised manual of the second edition of the Seashore's Measures of Musical Talents, the areas of application for the battery were defined as follows: “Educational and vocational counseling, admission to music instruction in schools, and selection for membership in bands and other music organizations” (Seashore, Lewis, & Saetveit, 1960, p. 3). Similarly, Wing (1962) noted that his test was “designed to pick out musically bright children at about the age of transfer to the secondary schools in order to give them the opportunity...of coaching in an orchestral instrument” (p. 39). The close interlocking of musical ability testing and placement practices is also evident from the work of Gordon (1965) and Karma (1983).

In contrast, the newer tests were generally devised to address research questions relating to the development, the correlates, and the determinants of musical ability. Accordingly, they are now used primarily in psychology and neuroscience. Reflecting this shift in aims and range of application, a greater emphasis is now placed on issues of stimulus design and control, psychometric soundness, as well as ease of use and accessibility. By these latter criteria, the earlier tests are clearly less compelling, and this is likely one of the reasons for their sparse use in contemporary research contexts.

At the same time, most of the earlier batteries were the result of many years of thought and research, the assumptions that guided their creation were well articulated, and contemporary measures draw from these earlier ones in several respects. Thus, we begin our review with a survey of four of the earlier musical aptitude tests, all of which gained a certain prominence and are sufficiently different from each other to warrant a separate entry. Subsequently, we will provide a survey of the newer generation of tests and conclude with a review of tests that were conceived for use in special populations.
Earlier Tests of Musical Ability

Seashore measures of musical talent  Seashore’s test was published in 1919, and a revised edition was released twenty years later (Seashore, 1919; 1939). The test manual to the second edition also underwent a revision (Seashore et al., 1960). By 1994 the test was out of print (Radocy & Boyle, 2012). Four characteristics of the Seashore’s approach stand out: (a) the notion that musical ability rests on the capacity for psychoacoustical discrimination; (b) the classification of psychoacoustical skills into several subcategories; (c) the measurement of these skills by means of pairs of pure tones, or brief sequences thereof; and d) the notion that in a true test of musical aptitude (rather than achievement) test scores are impervious to improvement by musical training.

To operationalize musical aptitude, Seashore created tasks in which listeners had to make certain discriminations, usually deciding whether two acoustical events were the same or different. For example, in the loudness subtest participants had to decide whether the second tone was stronger or weaker than the first tone. In the rhythm subtest, listeners were presented with tone patterns consisting of five to seven notes and had to determine whether the two patterns differed from each another. In the tonal memory subtest, participants were presented with a series of three to five tones that had no tonal relationship between one another. The participant’s task was to identify, by number, the tone that differed in the second compared to the first hearing. These examples may suffice to illustrate Seashore’s “atomistic” approach to assessing musical ability.

For many commentators it was difficult to see how the ability to discriminate between brief patterns of artificially sounding pure tones could be related to musical ability at large. To such criticism Seashore retorted that the aim of his battery was the measurement of aptitude not achievement, and that the use of elaborate or “holistic” musical material would inevitably result in the measurement of achievement (Seashore et al., 1960). To better understand this position, imagine an excerpt of Western tonal music being presented to a member of a professional orchestra and to a skilled Central African drummer. The particular choice of Western tonal music would inevitably confer an advantage to the orchestra player, due to her extensive knowledge of that music. In turn, the presentation of neutral tone sequences would give both participants about the same chance to succeed or fail. In Seashore’s logic, a superior performance by the drummer compared to the orchestra player would be indicative of the drummer’s superior musical aptitude despite the orchestra player’s obvious musical achievements, as evidenced in his or her status as member of a professional orchestra.

The example brings us back to the important conceptual distinction between achievement and aptitude, and the possibility that musical achievement may coexist with an absence of any notable musical aptitude or talent. As compelling as these views may appear, they must
ultimately be judged on empirical grounds. Seashore was aware of this requirement, and a number of validation studies were conducted. However, these studies have yielded mixed findings and are difficult to evaluate on the basis of the information provided in the relevant publications (e.g., Seashore et al., 1960; Shuter-Dyson & Gabriel, 1982). In part, the tenuous validity findings resulted from Seashore’s own misgivings about the very idea of validating musical ability tests:

They [the measure of musical talents] should not be validated in terms of their showing on an omnibus theory or blanket rating against all musical behavior, including such diverse and largely unrelated situations as composition, directing, voice, piano, violin, saxophone, theory, administration, or drums; because there are hundreds of other factors which help to determine job analysis in each of such fields...I have been bombarded all these years by the omnibusists for this type of validation, but I have persistently refused action on the ground that it had little or no significance. (Seashore et al., 1960, p. 384)

The Wing Standardized Tests of Musical Intelligence

Heralded as “undoubtedly the most important single contribution yet made to this branch of research” in the first review of the battery (Mainwaring, 1948, p. 290), the Standardized Tests of Musical Intelligence was devised by Herbert Wing, a student of intelligence researcher Cyril Burt. The test was primarily intended to help determine whether children should embark on a specialized musical education. The most salient feature of Wing’s test is that, in addition to assessing acuity of musical hearing along the lines of Seashore’s test, it also sought to capture aesthetic sensitivity to musical form and performance, what Wing referred to as musical appreciation. Accordingly, the second edition of the test was retitled “Tests of Musical Ability and Appreciation” (Wing, 1968).

Wing (1968) defines musical appreciation as “the power to recognize or evaluate artistic merit in music; it involves the deliberate aesthetic judgment of music as it actually exists in compositions rather than ability to solve problems connected with the elementary materials of which music is composed” (p. 2). The general term Wing used for both types of musical skills was musical capacity. To examine musical capacity, Wing devised tasks tapping into basic discrimination and tasks that required aesthetic judgment.

The test consists of seven subtests: chord analysis, pitch, memory, rhythmic accent, harmony, intensity and phrasing. The number of items per subtest range from fourteen to thirty. The stimuli are played on nine 10-inch records, and the length of the battery is about an hour. Three of the subtests are pure discrimination tasks (chord analysis, pitch, memory). For example, in the pitch subtest, participants were presented with a pair of chords of three to six notes. The second chord is either identical to the first chord, or had one note either raised or lowered by a semitone. Listeners are asked to mark “S” if the two chords are the same, “U” if the altered note moved up, and “D” if the altered note moved down (Wing, 1968, p. 50).
The other four subtests require an aesthetic judgment from the participant in addition to a perceptual one. Listeners are presented with two musical patterns and are first asked whether the samples are the same or different. If they notice a difference, the listener must decide which of the two is “better” or “more appropriate.” Thus, test 4 asks for the “better” rhythmic accent in two performances of the same piece. Test 5 requires the judgment of the more appropriate of two harmonizations of the same melody. Test 6 applies the same process to two ways of varying the loudness of a tune, and test 7 to two ways of phrasing the same piece of music. The total musical capacity score was the number of items answered correctly (Wing, 1968).

Information about the test’s reliability seems adequate (see table 16.1), but validity information is very sparse. In the second edition of the test only one unpublished examination of validity is briefly reviewed (Wing, 1968, p. 87). The full battery is very long (between sixty and ninety minutes) and quality of the original recordings is poor. These shortcomings prompted later researchers to recreate and use only the first three subtests of Wing’s battery (Delogu, Lampis, & Belardinelli, 2006; Schimikowski, Hemming, & Kleinen, 2003).

**Gordon’s tests of music audiation** Edwin Gordon developed a number of aptitude measures, two of which gained particular popularity: the Music Aptitude Profile (MAP, 1965), and the Advanced Measures of Music Audiation (AMMA, 1989). Very broadly, Gordon distinguished between musical achievement and aptitude, but felt that audiation was a preferable term to describe music aptitude (hence the use of music audiation in the AMMA). In Gordon’s (1965) own terms, “Audiation is 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)” (p. 8). Whereas audiation is defined as the capacity for processing music that is not physically present, aural perception refers to immediate processing of music. Generally speaking, audiation is the process of summarizing and generalizing from the specific music patterns we hear.

The general format of the tests is similar to those developed by Seashore and Wing in that listeners are asked to determine whether two music excerpts are the same or different. However, there is a greater similarity with Wing’s test in terms of the use of actual music instruments in the test excerpts as well as the inclusion of preference questions in addition to discrimination questions.

The MAP was developed for children in grades four through twelve and takes about forty minutes to complete. It was “to act as an objective aid in the evaluation of students’ music aptitudes so that the teacher can better provide for all students’ individual musical needs” (Gordon, 1995, p. 9). It consists of three main sections: tonal imagery (melody and harmony
<table>
<thead>
<tr>
<th>Test</th>
<th>Format</th>
<th>Sample</th>
<th>Subtests</th>
<th>Duration</th>
<th>Reliability</th>
<th>Validity (Pearson’s r coefficients)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Int. consistency&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Test-retest</td>
</tr>
<tr>
<td>SMMT (1919)</td>
<td>Offline, LP</td>
<td>C</td>
<td>6</td>
<td>60 min.</td>
<td>.55 to .84&lt;sup&gt;KR&lt;/sup&gt;</td>
<td>NR</td>
</tr>
<tr>
<td>STMI (1948)</td>
<td>Offline</td>
<td>C</td>
<td>7</td>
<td>60 min.</td>
<td>.91&lt;sup&gt;Split Half&lt;/sup&gt;</td>
<td>.76 to .88&lt;sup&gt;r&lt;/sup&gt;</td>
</tr>
<tr>
<td>MAP (1965)</td>
<td>Offline, CD</td>
<td>C</td>
<td>7</td>
<td>60 min.</td>
<td>.66 to .95&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.77&lt;sup&gt;r&lt;/sup&gt;</td>
</tr>
<tr>
<td>KMT (1973, 1975)</td>
<td>Offline, MP3</td>
<td>C &amp; A</td>
<td>1</td>
<td>20 min.</td>
<td>.66&lt;sup&gt;KR&lt;/sup&gt;</td>
<td>68&lt;sup&gt;r&lt;/sup&gt;</td>
</tr>
<tr>
<td>AMMA (1989)</td>
<td>Offline, CD</td>
<td>A</td>
<td>2</td>
<td>20 min.</td>
<td>.83 to .86&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.79 to .84&lt;sup&gt;r&lt;/sup&gt;</td>
</tr>
<tr>
<td>DTT (updated) (2001)</td>
<td>Online, CD</td>
<td>A</td>
<td>1</td>
<td>10 min.</td>
<td>NR</td>
<td>.77&lt;sup&gt;r&lt;/sup&gt;</td>
</tr>
<tr>
<td>MBEA (2003)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Offline</td>
<td>A</td>
<td>6</td>
<td>90 min.</td>
<td>NR</td>
<td>.75&lt;sup&gt;r&lt;/sup&gt;</td>
</tr>
<tr>
<td>Online Amusia Test (2008)</td>
<td>Online&lt;sup&gt;a&lt;/sup&gt;</td>
<td>A</td>
<td>3</td>
<td>15–30 min.</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>CAMP (2009)</td>
<td>Offline</td>
<td>A</td>
<td>3</td>
<td>37 min.</td>
<td>.69 to .92&lt;sup&gt;ICC&lt;/sup&gt;</td>
<td>.54 (CNC)</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Test</th>
<th>Format</th>
<th>Sample</th>
<th>Subtests</th>
<th>Duration</th>
<th>Reliability</th>
<th>Validity (Pearson’s r coefficients)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Int. consistency$^b$ Test-retest</td>
<td>Convergent</td>
</tr>
<tr>
<td>MET (2010)</td>
<td>Offline</td>
<td>A</td>
<td>2</td>
<td>20 min.</td>
<td>.94 to .96</td>
<td>NR</td>
</tr>
<tr>
<td>PROMS (2012)</td>
<td>Online</td>
<td>A</td>
<td>9</td>
<td>60 min.</td>
<td>.94$^a$; 95$^a$; 70 to .87$^b$</td>
<td>.33 to .68 (MAP, AMMA, MET, Vienna Symphonic Library)</td>
</tr>
<tr>
<td>Brief-PROMS (2012)</td>
<td>Online</td>
<td>A</td>
<td>4</td>
<td>25 min.</td>
<td>.70 to .87$^b$</td>
<td>.40 (harmonic closure task)</td>
</tr>
<tr>
<td>PROMS-Short (2017)</td>
<td>Online</td>
<td>A</td>
<td>8</td>
<td>25 min.</td>
<td>.92$^a$; .60 to .80$^b$</td>
<td>.58 (MMQ-Competence)$^d$</td>
</tr>
<tr>
<td>Mini-PROMS (2017)</td>
<td>Online</td>
<td>A</td>
<td>4</td>
<td>15 min.</td>
<td>.82$^a$; .53 to .76$^a$</td>
<td>.83$^b$; .51 to .69$^b$</td>
</tr>
<tr>
<td>GOLD-MSI (2014)</td>
<td>On/Offline</td>
<td>A</td>
<td>2</td>
<td>15 min.</td>
<td>.65 to .90$^a$</td>
<td>.60 to .70$^a$</td>
</tr>
</tbody>
</table>
SMDT (2014) Offline 3 10 min. .79 to .89

| Moderate (musicians vs. nonmusicians) Low to moderate hours of training |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| NR              | NR              | NR              |

The qualifications low, moderate, and high reflect Cohen's (1988) interpretational effect size guidelines \( r \approx .10; r \approx .30; r \approx .50 \), followed by the type of criterion in parentheses. When multiple studies were available an average was computed. The qualifications correspond to the reported size of the correlations and should not be taken as an overall judgment of the tests' criterion validity, which also depends on factors such as the adequacy of the selected criterion measures, and issues of sampling and generalizability.

a Online test is set up for online data collection.

b Single values refer to total test scores, value ranges to subtests.

c The qualifications low, moderate, and high reflect Cohen's (1988) interpretational effect size guidelines \( r \approx .10; r \approx .30; r \approx .50 \), followed by the type of criterion in parentheses. When multiple studies were available an average was computed. The qualifications correspond to the reported size of the correlations and should not be taken as an overall judgment of the tests' criterion validity, which also depends on factors such as the adequacy of the selected criterion measures, and issues of sampling and generalizability.

d MMQ = Mindedness Questionnaire (Zentner & Strauss, 2017). The MMQ is a self-report questionnaire that measures love of music (appreciation) and musical ability (competence).

e C = children; A = adults; NR = not reported; \( \alpha \) = Cronbach's alpha; \( r \) = Pearson's correlation; CNC = Consonant-Nucleus-Consonant test; SMDT = Swedish Music Discrimination Test.
discrimination); rhythm imagery (tempo and meter discrimination); and musical sensitivity (preference judgments of musical form and performance). In the preference subtests, the participant listens to two versions and two performances of musical excerpts and is asked to decide which of the two “sounds better.” The correct answer was determined by ten “reputable musicians,” at least nine of whom had to agree on which performance was the better one.

The AMMA is a music aptitude test for college students (both music majors and non-majors), high school students, and junior high students that lasts about twenty minutes. It consists of thirty trials, each of which includes a short musical “statement” followed after four seconds by a short musical “answer” with the same number of notes. The participant is asked to decide whether the statement and answer are the same or different. When the answer is different from the statement, the participant is also asked to decide whether the difference is a tonal or rhythm change. Within a certain test stimulus there may be either one or more tonal changes or one or more rhythmic changes, but not both, and the difference between the statement and the answer can occur at the beginning, in the middle, and/or at the end of the stimulus. Other aptitude measures developed by Gordon include the Intermediate Measures of Music Audiation (IMMA) for use in children between grades 1 and 6, and the Primary Measures of Music Audiation (PMMA), for use in preschool children up to grade 3.

The manuals of Gordon’s tests are excellent and they provide more detailed psychometric information than the manuals of earlier tests (see table 16.1). The validity information seems ample (e.g., Shuter-Dyson, 1982), but the picture is tempered by the fact that most validity examinations were carried out by Gordon himself or his students and published as booklets by GIA Publications—a publisher of sacred music and music education materials whose reviewing standards may not have been particularly stringent.

A closer look at some findings and test materials suggests that they may not have held up to careful scrutiny. For example, Gordon reported an association of $r = .97$ between MAP scores and teacher ratings of student’s musical talent (see Gordon, 1995, p. 85). Correcting for measurement error based on the MAP’s reliability, the result implies an impossible true correlation of $r > 1.00$. Some of the subtests raise questions about internal validity. For example, although the tempo subtest of the MAP intends to assess timing skills, it also heavily taxes tonal memory in that it not only uses melodic sequences, but also requires listeners to judge whether the tempo of the ending of the melodies is the same or different compared with the ending of the standard stimulus. As such, the MAP tempo test may be a better measure of tonal memory than of tempo skills per se (Law & Zentner, 2012).

**Karma Music Test (KMT)** The *Karma Music Test* (KMT) was developed by the Finnish music educator Kai Karma in the 1970s (Karma, 1973, 1975). Reflecting on the controversy between the “atomistic-psychoacoustical” approach adopted by Seashore and the “holistic-musical”
approaches advocated by Wing and Gordon, he describes the KMT as an attempt to strike a balance between these contrasting traditions. Thus, he notes that

the pendulum has swung from one extreme to another. Experts’ agreement in the preference items does not necessarily mean that an untrained child feels the same way; perhaps the “wrong” answers are just original and creative. … Familiarity with the music or the style used may give an unfair advantage to some subjects and make the test too much of culture and training. … This type of test is unlikely to find untrained potential; the question about why some are good and some not so good is practically left unanswered. (Karma, 1984, p. 28)

Karma objected that tasks included in Wing’s or Gordon’s tests came dangerously close to the tasks children learn to master as a part of their school education. What the test results basically reveal, then, is that “those who succeed in musical tasks will succeed in musical tasks” (Karma, 1984, p. 28). A true test of aptitude should predict performance on musical tasks without itself consisting of conventional musical tasks. According to Karma, the core capacity underlying various culture-specific expressions of musical ability is “the ability to conceive the structure of acoustic material” (Karma, 1973, p. 9). Just as spatial ability can be understood as an ability to conceive visual patterns, musical ability can be defined as the ability to perceive auditory patterns or sets of relations between tones. This ability can be assessed as early as “after one school year” (Karma, 1984, p. 29). With time, this general ability would ramify into the domain-specific skills that we associate with particular musical systems, such as the understanding of tonality in traditional Western tonal music.

To measure this general ability for auditory structuring Karma created two to six simple tone sequences, which are repeated without pauses or other indications where one sequences ends and the other begins. Participants are presented with a reference and a comparison pattern that are either the same or different. A correct answer is seen as evidence that the subject has identified the pattern. The difficulty of the tone sequences is manipulated by placing the stress on more or less intuitive parts within the sequence. The most recent version of the test includes forty items (Karma, 2007).

Like Seashore, Wing, and Gordon, Karma was aware of the challenges involved in establishing test validity for a music aptitude test. His views of the subject are too elaborate to be reproduced here (but see Karma, 2007, for a summary of key concerns). In one study, test scores in over three hundred children under age ten were found to correlate between $r = .05$ and $r = .25$ with the children’s previous amount of musical instruction and between $r = .32$ and $r = .35$ with expert ratings of the children’s performance on entrance examinations of a Finnish music school (unpublished, reported in Karma, 2007, pp. 88–89). The test is widely used “as one selection instrument in nearly half of the music schools in Finland as well as music oriented classes” (Karma, 2007, p. 85) and has also been used in research
examining the molecular genetics of musicality (see the section on “Determinants of Individual Differences,” below).

**Recent Musical Ability Tests**

In the review of newer tests that follows, coverage is restricted to tests that have been published after the year 2000 in English-language, peer-reviewed professional journals. This might leave out tests published in other formats and languages. Our review begins with the *Profile of Music Perception Skills* (PROMS, Law & Zentner, 2012), because it is the most inclusive of the newer tests. The musical skills measured by the PROMS are also included in other recent tests, whereas other skills assessed by the PROMS (e.g., timbre, tuning, tempo, accent) are unique to the PROMS. Starting with the PROMS therefore allows for a more economic presentation of the various tests.

*The Profile of Music Perception Skills (PROMS)* Work on the PROMS began in 2009, when many of the musical ability tests reviewed earlier were either difficult to access or limited in one or more essential psychometric features. One extensively validated battery, the Montreal Battery Evaluation of Amusia (discussed further below), was developed to assess amusia rather than variations in musical abilities within the normal range. This gap motivated Law and Zentner (2012) to create a timely, easily accessible musical ability test to validate in adult populations.

The key question in the initial stages of test-development was whether this instrument should assess (a) the comprehension of a specific, culturally evolved musical system or (b) the ease of processing for basic patterns of rhythm and sound that can be found across various musical systems and traditions. The authors chose the latter variant for reasons that are consistent with some of the concerns expressed by earlier test authors. First, an almost endless spectrum of musical varieties has been described in the ethnomusicological literature (e.g., Nettl, 2015). It is practically impossible to incorporate such variety into a battery. Moreover, choosing any particular type of music would confer an advantage to listeners who are familiar with the chosen type of music. This bias can be largely avoided through the use of basic tone patterns that are relatively neutral stylistically. Second, small units of music are suitable for testing and can be used for musically trained and untrained populations alike. Third, they offer greater possibilities for stimulus control to the test-developer.

With these basic considerations in mind, the authors set out to create a battery that should meet four criteria: (1) the test should be equally suitable for listeners with different musical backgrounds, including no musical background; (2) the test should be more inclusive than previous batteries with respect to the musical perceptual components tested; (3) the test should assess each perceptual component with the greatest possible specificity; and (4) the
test should meet contemporary standards for test construction in terms of validity and reliability (Law & Zentner, 2012). To this end, the authors created tasks that, without being as reductionist as those devised by Seashore or Karma, do not strongly rely on traditional Western tonal music.

A particular emphasis was devoted to the development of tasks that were almost completely absent from previous batteries and that would tap into sensitivity for timbre and for the purity of harmonic (as opposed to melodic) intervals. The result was a battery assessing perceptual skills across eight domains: melody, tuning, tempo, rhythm, stress/accent, pitch, timbre, and loudness. Because a complete description of the PROMS has been provided elsewhere (Law & Zentner, 2012), here we recapitulate only some of its most characteristic features.

In each of the subtests, participants are required to judge whether a reference and a probe stimulus are the same or not. There are eighteen comparisons, or trials, per subtest, with an equal number of same-correct and different-correct answers. The authors also introduced the Brief PROMS, a version that consists of only four of the nine subtests: melody, tuning, tempo, and accent. The entire version (henceforth Full PROMS) takes about one hour to complete, the Brief PROMS around twenty-five minutes. In each subtest the trials range from easy to difficult. Difficulty level was manipulated in different ways, depending on the subtest.

For example, in the melody subtest participants hear a two-bar monophonic harpsichord melody twice, followed by the probe melody, which can differ by one or more tones. The accent subtest assesses skills in discerning the relative emphasis given to certain notes in a rhythmic pattern, which is comparable to recognizing stress in speech. In the accented notes the intensity was raised by 3 decibels (dB). In the easy test trials, intensity changes were applied to most sound events so as to increase the probability of detecting the alteration. In the more difficult trials changes in accentuation were less frequent and thus required more subtle perceptual skills to be detected (see figure 16.1).

In the timbre subtest, original instrumental sounds are used to produce 4-note chords (C4, E4, G4, C5) in various instrumental configurations. The difficulty is manipulated by means of subtle changes to the instrumentation in each chord. In easy trials, the comparison consists of reference and probe chords being played by different families of instruments, such as horn versus strings. In more difficult trials, the probe and reference chord are the same, except for one instrumental part, which is taken from a different instrumental group (e.g., four woodwind parts vs. three woodwind and one string part). In the most difficult trials, the replacement occurs within the same instrumental family (e.g., four viola sounds vs. three viola and one violin sound; see figure 16.2). In the tuning subtest, the difficulty level of the test trials was varied by subtle manipulations to the E note in the chord, with a range from 50 to 10 cents, as is illustrated in figure 16.3.
Figure 16.1
Illustration of the PROMS subtest “accent.” The top figure shows the intensity domain of the accent subtest, and the bottom figure shows the time domain of the accent subtest. As the top figure shows, the intensities of the accent notes (a) are represented by in the time domain figures, Accent (a’). Accent (b) shows that the unaccented notes (second, third, and fourth beats) are 23 dB lower than the accented note, which can also be seen in the comparison-stimulus in the time domain, Accent (b’). The example of a complex trial shows the alteration affecting only one or two events.
*Represents the alteration in the comparison-stimuli.
Figure 16.2
Illustration of the PROMS subtest “timbre.” The easy trial consists of two groups of instruments from altogether different families. In the complex trial, the instrument changes on only one note are taken from the same family (strings).

Figure 16.3
Illustration of the PROMS subtest “tuning.” The difficulty of tuning trials is manipulated by the extent to which the note E4 is shifted out of its proper frequency (from 10 to 50 cents).
Taking the need for brevity into consideration, two shorter versions of the Full and the Brief PROMS were developed more recently: the PROMS-Short, or PROMS-S, and the Mini-PROMS (Zentner & Strauß, 2017). Based on a psychometric analysis of data collected after the initial test publication and the elimination of one of the subtests (loudness), the number of trials was halved, resulting in test-durations of about twenty-five minutes for the PROMS-S and around fifteen minutes for the Mini-PROMS.

Internal consistency and test-retest reliabilities for all versions of the PROMS exceed .85. The values for the individual subtests are lower but acceptable given the relatively small number of trials per subtest (see table 16.1). Convergent validity was established with the relevant dimensions of Gordon’s Advanced Measures of Music Audiation and Musical Aptitude Profile (melody, rhythm, tempo), the Musical Ear Test (rhythm), as well as sample instrumental sounds from the Vienna Symphony Orchestra library (timbre).

Criterion validity was evidenced by consistently significant and sizeable relationships between test performance and external musical proficiency indicators, notably self-reported years of musical training, involvement in critical listening activities, music degrees and qualifications, and musicianship status (rs = .37 to .63). The correlation with the composite score across the four dimensions was $r = .57$, $p < .01$ (see table 16.1). A separate group of investigators examined the criterion validity of the Brief-PROMS against performance on a harmonic closure task (Kunert, Willems, & Hagoort, 2016). The association between the PROMS full score and the harmonic task performance was $r = .40$ ($p < .01$). Because the PROMS does not assess knowledge of traditional Western harmony, the correlation is nontrivial, and provides some evidence that the skills assessed with PROMS might generalize to other musical skills that are not measured with the test.

Discriminant validity was examined against the gap detection task with white noise (Law & Zentner, 2012), against IQ (Zentner & Strauß, 2017), and against short-term and working memory by an independent group of researchers (Kunert et al., 2016). The gap detection task is an established test to measure nonmusical auditory sensitivity test that was chosen because it does not have a strong pitch component. Correlations between the gap detection task and the PROMS total and subtest scores were small and nonsignificant ($r's < .20$, see Law & Zentner, 2012; see also table 16.1). In addition, Kunert et al. (2016) found good discriminant validity with short-term memory and working memory (as assessed by the Wechsler Intelligence Scale). Thus, backward digit span correlated neither with the PROMS total score nor with any of its subscales ($r < .10$). The correlation between forward digit span and the PROMS total score was slightly larger ($r = .22$), but smaller than the correlations that have been reported with other music ability tests (see table 16.1).

A convenient new feature called Modular PROMS offers researchers the possibility to request customized batteries that may include any combination of the subtests. Apart from
the flexibility in content, this feature also allows researchers to vary the duration of the test from as little as four to five minutes (if only one subtest is included) to between twenty-five and thirty minutes (if all subtests of the PROMS-S are included). Furthermore, the instruments are administered and scored completely online without the need for an administrator. As such, they can be administered in a school computer lab, under individual testing conditions in a laboratory, or on adult participants’ personal computers outside the laboratory. Researchers who are interested in using the test for research purposes receive a unique URL that provides access to their own uses of the PROMS. Questions or questionnaires that are of interest to researchers can be easily added to the musical battery. The data are securely stored on a university server and the data collected can be downloaded at any time as an Excel or an SPSS file. Finally, the new tools are available in several languages—at present in English, French, German, Italian, Japanese, Norwegian, Russian, and Spanish (Zentner & Strauß, 2017).

**The Musical Ear Test (MET)**

The Musical Ear Test (MET) was developed by a Danish team led by Mikael Wallentin (Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010). The authors note that Gordon’s AMMA is “the test that comes closest to our new test” (Wallentin et al., 2010, p. 189). However, whereas rhythmic and melodic changes are embedded in the same melodic context in the AMMA, the MET tests melody and rhythm separately.

The test is based on same-different judgments of short melodies (three to eight tones) played with sampled piano sounds and of rhythmic patterns (four to eleven beats) played with woodblocks sounds. There are fifty-two melody trials and fifty-two rhythm trials, with half of the trials being the same and half being different. The stimuli, which can be obtained by contacting the first author, consist of MP3 and WAV files. Answer sheets and correct answers are provided as Excel files. Both subtests exhibit a high internal consistency, with Cronbach alpha values of .96 for the melody subtest and .94 for the rhythm subtest. The test takes about twenty minutes to complete.

In terms of validity, MET scores have been found to distinguish significantly between nonmusicians, amateurs, and professional musicians, and there are no ceiling effects (even professional musicians rarely score 100 percent). Moreover, MET scores correlate highly with results obtained on an imitation test used in Danish music academies ($r = .89$, $p < .001$), in which participants reproduce short rhythmical and melodic phrases by clapping hands or singing (see table 16.1).

**The Swedish Musical Discrimination Test (SMDT)**

The Swedish Musical Discrimination Test (SMDT) (Ullén, Moring, Holm, Eriksson, & Madison, 2014) was developed to provide measures of basic aspects of musical ability operationalized by means of discrimination tasks. The test includes three subtests measuring discrimination skills for short melodies, rhythmic patterns, and single pitches. The authors do not explain the rationale that guided these
choices, but note that the subtests are similar to the Tunes, Rhythm, and Pitch tests of the Bentley measures of musical ability (Bentley, 1966) and to the corresponding subscales of the MET and PROMS tests. More generally, the SMDT is described as following the “atomistic” tradition of Seashore, as opposed to the “omnibus” approach of Wing. An Adobe Flash Player implementation of the test can be obtained by contacting the first author.

The melodies (four to nine tones) consist of eighteen pairs of isochronous sequences of sampled piano tones for which the pitch of one tone differed. Participants had to identify the tone that differed between both melodies. The rhythmic patterns consist of eighteen pairs of sequences of brief sine tones (five to seven tones), and participants have to judge whether the sequences in each pair are the same or different. Finally, the pitch task consists of twenty-seven pairs of sine tones, and participants are asked to identify whether the first tone in a pair was higher or lower in pitch than the second one.

Each subtest takes about four minutes to complete and can be administered online. Scores on the three subtests are moderately correlated with each other (r coefficients between .28 and .41 for all three subtests, all ps < 0.001). Internal consistency values (Cronbach alpha) range between .79 and .89 for the three subtests. As with the MET scores, individuals having taken music lessons scored significantly higher on average than individuals with no music training (Cohen’s $d$, a measure of effect size corresponding to the difference between means divided by the standard deviation, yielded values between .35 and .60 for all three subtests), and total hours of music training were somewhat correlated with SMDT scores for participants who had played an instrument (see table 16.1).

**Goldsmiths Musical Sophistication Index (Gold-MSI)**  
The Goldsmiths Musical Sophistication Index (Gold-MSI) is comprised of both a self-report questionnaire and a test battery (Müllensiefen, Gingras, Musil, & Stewart, 2014). The test battery involves same-different judgments in a melodic memory task consisting of thirteen short pairs of melodies (based on Bartlett & Dowling, 1980) and a beat tracking task (based on Iversen & Patel, 2008), in which participants have to judge whether the beat track matches the musical beat in seventeen excerpts. The completion of the test battery requires approximately fifteen minutes.

Stimuli for both tasks (WAV files), as well as population norms and scoring templates (in Excel format), can be downloaded on a website administered by the first author, and a Psychopy (Peirce, 2007) implementation of the test battery is available. The test battery is available in both English and German, and validated versions of the questionnaire part of the Gold-MSI are available in English (Müllensiefen et al., 2014) and German (Schaal, Bauer, & Müllensiefen, 2014; Fiedler & Müllensiefen, 2015).

Internal consistency values (Cronbach alpha) reached .65 for the melodic memory task and .90 for the beat tracking task, and test-retest reliabilities ranged between .60 and .70 (see table 16.1). Scores obtained on both tasks were significantly correlated with the scores
obtained on the self-report questionnaire \( (r = .37, p < .001) \) and with the amount of musical training, as determined by a random forest regression model. Gold-MSI scores have also been correlated with other test batteries and with personality measures such as the Big Five Inventory (Greenberg, Müllensiefen, Lamb, & Rentfrow, 2015) (see also table 16.1).

Furthermore, both the SMDT and the Gold-MSI have been validated in online studies with samples consisting of several thousand individuals (two subsamples of ~3,400 individuals for the SMDT and over 140,000 participants for the Gold-MSI), which suggests that both tests are suitable for online administration to large and sociodemographically diverse samples (Müllensiefen et al., 2014; Ullén et al., 2014). For the Gold-MSI, the correlation between self-reported musical sophistication and performance on the melody task was somewhat lower in the large online sample than in the test-retest sample of thirty-four laboratory-tested participants \( (r = .29 \text{ vs. } r = .51) \), but this decrease in effect size between controlled laboratory conditions and online administration is commonly observed (Birnbaum, 2004).

Tests for Special Populations

The Distorted Tunes Test (DTT) The *Distorted Tunes Test* (DTT) was developed by Fry in the 1940s to study the prevalence of “tone deafness” or dysmelodia, corresponding to an inability to detect pitch changes in melodies (nowadays known as amusia), in the British population (Fry, 1948). It consisted of twenty-six short popular melodies ranging in length from twelve to sixteen notes. Sixteen of the melodies were distorted by a pitch change of two to nine notes, generally within one or two semitones of the correct note. Rhythm and contour were unaltered. Subjects were asked whether the melody was correct or incorrect, and whether they were familiar with the melody. Studies using the DTT showed that cultural biases could be reduced by carefully selecting melodies, and that test-retest scores were stable over decades. Interestingly, around 5 percent of the population scored no better than chance on the DTT. Drayna, Manichaikul, de Lange, Snieder, and Spector (2001) proposed an updated version of the DTT, retaining eight melodies from the original DTT and adding eighteen new ones.

The updated version, which follows the same construction principles as the original one, was validated in American and British populations. As with the original DTT, test-retest scores were highly correlated \( (r = 0.77) \). To assess the influence of long-term musical memory on performance on the DTT, Drayna et al. (2001) also developed an *International Tunes Test* (ITT), consisting of eighteen unfamiliar melodies that used both Western and non-Western tonal systems. Scores on the ITT were highly correlated with DTT scores \( (r = 0.71) \), suggesting that long-term musical memory is only a marginal predictor of DTT scores.¹

The Montreal Battery for the Evaluation of Amusia (MBEA) A modern analog to the DTT, the *Montreal Battery for the Evaluation of Amusia* (MBEA) was initially developed to assess musical...
abilities of brain-damaged patients who suffer from acquired amusia, but is nowadays used to diagnose both acquired and congenital forms of amusia. Based on a model of music processing described in Peretz and Coltheart (2003), the MBEA consists of six subtests, three of which test melodic organization (Scale, Contour, and Interval subtests), two test temporal organization (Rhythm and Meter subtests), and one tests melodic memory (Memory subtest). Unlike the DTT, which uses popular melodies, the MBEA controls for familiarity effects by employing, for all subtests, a selection of novel musical phrases composed according to the principles of the Western tonal system. These phrases are monophonic—that is, they consist of a single voice for all but the metric test, in which the melodies are polyphonic and twice as long. The participants are presented with two practice trials and thirty-one experimental trials. A trial consists of a reference melody and a comparison melody, which are separated by a 2-second silent interval. Fifteen of the comparison melodies are identical to the reference melody, and fifteen have comparison melodies that differ with respect to the reference melody. In the Scale subtest, the altered melodies violate the key but keep the overall contour intact; in the Contour subtest, they violate the contour while keeping the key intact; and in the Interval subtest, key and contour are kept intact but the pitch interval is violated. For the Rhythm subtest, the rhythmic grouping of the comparison melody is changed by altering the duration of two adjacent notes. In addition to those thirty trials, each subtest contains a catch trial to ensure that the participants are paying attention and not simply guessing.

Whereas in the first four subtests (Scale, Contour, Interval, and Rhythm), listeners are presented with pairs of melodies and asked to judge whether the two melodies are the same or different, in the last two subtests (Meter and Memory), listeners are presented with a single melody on each trial. For the Meter subtest, listeners judge whether the presented melody is a march or waltz, whereas for the Memory subtest, listeners judge whether the melody is one that they heard on the previous subtests (an old melody) or is previously unheard (a new melody).

Peretz, Champod, and Hyde (2003) report that, based on a retest of twenty-eight participants four months after initial testing the MBEA displays test-retest reliability. Convergent validity of the MBEA was examined with two subtests (melody and meter) of the Musical Aptitude Profile. These two subtests, which were chosen because they were closest in content and format to the MBEA, were administered to sixty-eight subjects. Scores on the two batteries were positively correlated ($r = 0.53, p < 0.01$) (see table 16.1). A closely related test for the online identification of congenital amusia was also developed to facilitate the screening of potential cases of amusia in the general population (Peretz et al., 2008). This test, which is comparable to the Scale and Meter subtests of the MBEA, consists of three conditions with a total of seventy-two melodies derived from twelve stimuli from the MBEA. Scores on the online test are highly correlated with scores obtained on the full MBEA ($r = .82$), indicating good convergent validity with the parent instrument.
A widely used measure of amusia, the MBEA has recently been criticized for its reliance on percent-correct instead of $d'$ scores (Henry & McAuley, 2013; Pfeifer & Hamann, 2015). The advantage of $d'$ scores is that they correct for participants' response bias, thereby reducing the likelihood of people being diagnosed as amusics due to response bias rather than their musical discrimination ability.

**Clinical Assessment of Music Perception Test (CAMP)** The University of Washington Clinical Assessment of Music Perception Test (CAMP) is a computerized test that assesses pitch direction discrimination, melody recognition, and timbre recognition ability of cochlear implant patients (Kang et al., 2009). The pitch and melody subtests use a digitally synthesized synthetic piano sound. Listener scores on these subtests are based on the percentage of correct answers. In the pitch subtest, each listener is presented with two tones differing in pitch by one to twelve semitones, and is asked to indicate whether the first or second note had the higher pitch. This subtest uses a two-alternative forced, 1-up 1-down adaptive testing method (Levitt, 1971). The melody subtest consists of twelve commonly known melodies played in an isochronous manner. Each melody is played three times (with melodies presented in a random order), and listeners are asked to identify the melodies by selecting the respective title from the answer choice. The timbre subtest consists of eight instruments of four major instrument classes (strings, brass, woodwinds, and percussion).

All instruments are played in an identical sequence of five notes: C4-A4-F4-G4-C5. As in the melody test, each melody is played three times, and listeners are asked to identify the timbre by selecting the labeled icon of the instruments from the given answer choice. Test-retest reliability (intraclass coefficients) were .85, .92, and .69 for pitch, melody, and timbre subtests, respectively. The melody, timbre, and pitch subtests were shown to correlate moderately with two widely used speech perception tests for Cochlear implant users, the Consonant Nucleus Consonant test CNC and the Speech Recognition Thresholds (SRTs) (see table 16.1).

**Determinants of Individual Differences in Musical Ability**

Most test batteries for the assessment of musical ability were devised to detect untrained musical potential rather than musical accomplishment. As such, they were based on the assumption of individual differences in innate musical giftedness. The challenges involved in separating innate talent from practice-based accomplishment were well known to the pioneers of musical ability testing, and questions about the determinants of musical ability have been a subject of debate since the days of Seashore. However, because methods for examining the relative contribution of nature and nurture to individual differences in musical ability were comparatively primitive during the early days of musically testing, initial views on
the subject were based on speculation rather than scientific evidence. Over the past two
decades, our knowledge of the genetic and environmental contribution to musical ability has
expanded considerably. In what follows, we review some of this evidence in terms of genetic
contributions to musical aptitude, the role played by biological and environmental factors,
and the factors influencing the emergence of true musical expertise from raw musical talent.

Genetic Contributions to Musical Aptitude
The administration of musical aptitude tests on a large scale has shown a wide range of
variation in musical aptitude in the general population (Müllensiefen et al., 2014). Analogously,
genetic diversity in contemporary human populations has recently been shown to
be greater and more wide-ranging than previously thought (Abecasis et al., 2012), which, in
turn, reinforces the possibility of a link between interindividual variation at the genotypic
level and phenotypic (behavioral) variation in the music domain. Thanks to modern genetic
techniques, researchers are now in a position to examine this link empirically (Gingras,
Honing, Peretz, Trainor, & Fisher, 2015). The relationship between phenotype and genotype
is typically investigated through behavioral genetic approaches such as familial aggregation
analysis, which evaluates whether a trait tends to cluster in families above chance level, or
twin studies, which disentangle the contribution of genetic and environmental factors by
comparing the similarity of a phenotype in monozygotic (identical) versus dizygotic (non-
identical) twins. Complementing these approaches, molecular genetic methods such as linkage
analysis, association analysis, and copy number variation analysis are used to attempt to
map the genetic loci thought to be associated with the trait under study (for reviews on the
genetic basis of music aptitude, see Gingras et al, 2015; Tan, McPherson, Peretz, Berkovic, &
Wilson, 2014).

From a genetic standpoint, pitch perception has been examined more extensively than
other facets, although a number of recent studies have investigated the heritability of other
facets such as rhythm perception. One of the earliest twin studies on music perception abili-
ties in the general population was conducted by Drayna and colleagues (2001), who admin-
istered the DTT (discussed above) to 136 identical and 148 nonidentical twin pairs. Scores
on the DTT, which can be considered a proxy for the participants’ ability to judge successive
pitch intervals, were found to have a heritability of 71–80 percent, with no significant con-
tribution of the shared environment; this result indicates a strong hereditary influence on
pitch discrimination abilities.

A recent study assessed the melodic and rhythmic perception abilities of 69 identical and
44 nonidentical twin pairs, as well as 158 individual twins (Seesjärvi, Särkämö, Vuoksimaa, Ter-
vaniemi, Peretz, & Kaprio, 2016), using the MBEA test (discussed above). Here, 58 percent of the
variance on a same-different judgment task comparing two melodies differing by one note was
attributable to additive genetic effects, in line with earlier findings by Drayna et al. (2001). On the other hand, for an “in-key/out-of-key” detection task and for an “on-beat/off-beat” rhythm detection task, most of the variance was accounted for by environmental effects, which suggests a lesser role for heritable factors.

Research examining the genetic correlates of musical memory yielded results suggesting a role for the arginine vasopressin receptor 1a (AVPR1A) and serotonin transporter (SLC6A4) genes (Granot et al., 2007), an interpretation that was reinforced by a follow-up experiment showing that intranasal administration of the arginine vasopressin hormone impaired musical working memory, whereas mood and alertness levels were enhanced (Granot, Uzefovsky, Bogopolsky, & Ebstein, 2013). Interestingly, arginine vasopressin and its receptor are thought to play a far-reaching role in social attachment behaviors in rodents and humans (Insel, 2010).

A series of studies in a sample of extended Finnish families has contributed to the characterization of the importance of hereditary factors to musical aptitudes (Oikkonen et al., 2015; Pulli et al., 2008; Ukkola-Vuoti et al., 2011; Ukkola, Onkamo, Raijas, Karma, & Jarvela, 2009). The first of these studies (Pulli et al., 2008) tested fifteen families (234 people) on a series of music perception tests comprising the Karma Music Test and Seashore’s pitch and rhythm subtests (both discussed above). The analysis yielded heritability estimates of 42 percent for the Karma Music Test, 57 percent for Seashore’s pitch subtest, and 21 percent for Seashore’s rhythm subtest. These results are consistent with the findings of Seesjärvi et al. reported above (obtained with a different sample and methodology), with both studies indicating a higher heritability for pitch perception than for rhythm perception.

Another study by the same Finnish group investigated whether previously identified candidate genes, such as AVPR1A and SLC6A4 (Granot et al., 2007), were associated with musical aptitude, as measured by the Karma and Seashore tests (Ukkola et al., 2009). Certain haplotypes of AVPR1A were found to be associated with music perception abilities, but no other candidate genes emerged. A follow-up study evaluating the music listening activities of thirty-one Finnish families (437 members) yielded an association between AVPR1A and active music listening (Ukkola-Vuoti et al., 2011).

The same research team then conducted a genome-wide analysis of 767 individuals from seventy-six families, whose music perception abilities were once again evaluated using the Karma and Seashore tests (Oikkonen et al., 2015). The strongest associations were observed in the vicinity of the GATA binding protein 2 (GATA2) gene, which regulates the development of cochlear hair cells and the inferior colliculus, both implicated in tonotopic mapping. However, while the study suggests a promising link between music perception abilities and auditory pathway genes, it did not support previous findings by the same group on candidate genes such as AVPR1A.
A few large-scale twin studies have focused on music production abilities. Coon and
Carey (1989) analyzed music-related data obtained from an earlier survey which included
personality and interest questionnaires. They found higher heritability for participation in
singing activities than for self-reported musical aptitudes, and heritability estimates were
lower for females than for males. Using self-reported data from 1,685 twin pairs (twelve- to
twenty-four-year old), Vinkhuyzen, van der Sluis, Posthuma, and Boomsma (2009) esti-
mated the heritability of aptitude and exceptional talent across various domains such as
language, mathematics, and sports, as well as music. Again, heritability for musical apti-
tude was lower for females (30 percent) than for males (66 percent). Whereas these two
studies did not objectively assess musical abilities, a more recent study compared the heri-
tability of self-reported musical achievement with objectively evaluated musical aptitude
(using the SMDT test, discussed above) in a large sample of 10,975 Swedish twins (Mosing
et al., 2015).

Once more, the heritability of music achievement was lower for females (9 percent) than
for males (57 percent). However, heritability estimates for musical aptitude test scores were
51 percent for females and 38 percent for males, reversing the trend observed when using
self-reports. Another study on the same set of twins identified two shared genetic factors
explaining the covariation between musical aptitude and intelligence, one explaining the
covariation between the SMDT scores and IQ (with between 32 and 49 percent of the vari-
cance on the SMDT scores explained by this shared factor), and a second one accounting for
the variance shared specifically among the various SMDT subtests (with 15 percent of the
variance for pitch, 20 percent for rhythm, and 64 percent for melody due to this factor).
These results suggest that the positive correlation between musical aptitude and intelligence
stems from broad genetic influences on cognition, and possibly from genes specifically con-
trolling auditory functions (Mosing et al., 2014).

Research exploring genetic contributions to music production abilities has largely focused
on singing abilities, which is probably the most widespread musical activity in the general
population. Morley and colleagues (2012) tested for the association of AVPR1A and SLC6A4
polymorphisms that were previously associated with musical abilities (Granot et al., 2007;
Ukkola et al., 2009) (but see Oikkonen et al., 2015, for nonreplication) and creative dancing
(Bachner-Melman et al., 2005) with choir participation in 523 subjects (Morley et al., 2012).
Results showed a significant association for a SLC6A4 polymorphism but not for the AVPR1A
haplotypes identified in previous studies.

Park and colleagues (2012) investigated the genetic determinants of scores obtained on a
vocal pitch-production accuracy test, identifying a region on chromosome 4 that overlaps
with a region identified by the Finnish group (Pulli et al., 2008; Oikkonen et al., 2015). More
specifically, a region near the gene UGT8, a gene coding for an enzyme playing an important
role in the synthesis of a component of myelin membranes in the central nervous system, was significantly associated with performance on the pitch production task.

Clearly, the feasibility of large-scale genetic studies of musical ability depends critically on the ability to obtain a robust, objective, and reliable measure of musical aptitude. Thus, the development of short online tests that capture a broad range of musical abilities and are applicable to the general population is a major step forward. Although many of the studies discussed in this section have used self-reports (e.g., musical creativity studies [Ukkula-Vuoti et al., 2013; Ukkola et al., 2009], twin studies on music ability [Coon & Carey, 1989, Vinkhuyzen et al., 2009], and participation in singing activities [Coon & Carey, 1989]), more recent studies have made use of large-scale online testing (e.g., Mosing et al., 2014).

Biological and Environmental Factors
Considering that parents share 50 percent of their genetic material with their children, environmental and genetic effects must be considered together when evaluating familial influences over the child’s musical development. According to Scarr & McCartney (1983), familial influences can be categorized in three main types of nature-nurture interactions. In passive interactions, the child’s musical behavior is influenced through interactions with the parents. Thus, children of musically active parents are more likely to be exposed to, and participate in, musical activities, and such an environment is likely to correspond to the child’s genetic potential, given the child’s genetic kinship with his or her parents. Evocative interactions describe contexts in which parents respond to and support behaviors evidencing musical talent, such as encouraging an obviously gifted child to pursue musical training, whereas active nature-nurture interactions occurs when the child selects and shapes his or her own environment according to the child’s musical interests.

Familial atmosphere also plays a role in the development of musical talent. Csikszentmihalyi, Rathunde, and Whalen (1993) used a questionnaire to evaluate familial atmosphere and found two main axes, labeled integration (emphasis on support and harmony) and differentiation (emphasis on involvement and freedom). This led to a classification of families into four main types: simple (low integration and differentiation), integrated (high integration and low differentiation), differentiated (low integration and high differentiation), and complex (high integration and differentiation). Results indicated that “complex” families provided the optimal environment for nurturing and developing talent. Further corroborating evidence pointing to the importance of the familial environment was provided by Bastian (1992), who showed that musical giftedness was better predicted by the income and amount of musical activity of the parents than by the test scores or performance achievements of young children. Interestingly, there is a significant relationship between the children’s interest in music and the musical activity of the father, but not that of the mother (Vogl, 1993).
It should be noted that the array of skills necessary to acquire expertise differs according to the musical genre: Western classical music prioritizes technique and exactitude, whereas other genres such as jazz value “playing by ear” to a larger extent (Kleen, 1997; Woody and Lehmann, 2010). Even within one musical style such as Western classical music, the skills required for a composer and for an instrumentalist diverge substantially (Simonton, 1999; Oerter, 2003).

Besides genetic and environmental factors, other biological variables, which are often influenced both by genes and environment, play a role in the development of musical aptitude. Thus, some physical attributes are associated with expertise in playing specific families of instruments. For instance, possessing broad left-hand fingertips and large hands (especially for cello and bass) seems to be advantageous for string players, whereas even teeth and a large hand span may help master wind instruments (Mills, 1985). Depending on the particular instrument, the physical attributes may be more specific: for instance, whereas thin lips are helpful for flutists and trumpet players, thick lips are an asset for trombonists and tuba players.

Musical proficiency also seems to be correlated, at least to some extent, with hormonal levels. Thus, musicians show higher levels of melatonin than nonmusicians (Hassler & Gupta, 1993). Melatonin is involved in several regulatory systems, including the circadian rhythm and the immune system. Hassler (1992) compared salivary testosterone levels and spatial ability of musicians and nonmusicians, and found that an optimal testosterone range appears to be in the lower normal range for males and in the upper range for females, a result that points to androgyny. Moreover, according to Hassler (2000), these hormonal differences between musicians and nonmusicians are already present in utero, which suggests that there are innate biological differences between the two groups. Additionally, musicians were found to obtain significantly higher scores on spatial ability tests than nonmusicians (Hassler, 1992). However, in contrast to Hassler (1992), Sluming & Manning (2000) found that male elite musicians had a lower ratio of the length of the second and fourth digits (2D:4D ratio), which is associated with high testosterone levels. Moreover, the higher the rank of the musicians in a British symphony orchestra, the lower the 2D:4D ratio was. No differences in 2D:4D ratio were found among instrument groups, which suggests that the ratio was not related to other attributes that may be advantageous for playing specific instruments.

Last but not least, musical training induces measurable changes in the brain anatomy. Hyde et al. (2009) demonstrated that musical training shapes structural brain development. After fifteen months of musical training in early childhood, structural brain changes were observed in the motor and primary auditory cortices. These structural changes were correlated with improvements in musically relevant motor and auditory skills. However, musical training did not produce significant improvement in nonmusical cognitive skills. For a more
detailed treatment of this topic, the interested reader should consult chapters 7 and/or 10 in this volume.

**Nurturing Talent: From Ability to Expertise**

Clearly, talent is not sufficient to become a professional musician. Training and practice are necessary to develop the cognitive and motor mechanisms for high-level musical skills, and learning involves cognitive and neurophysiological changes. Indeed, after reviewing findings showing that individual differences in basic capacities and abilities are poor predictors of performance, some authors claim that the influence of innate factors is small or “possibly even negligible” (Ericsson & Lehmann, 1996). In support of this view, it has been found that cognitive ability was not a significant predictor of piano-playing skills (Krampe & Ericsson, 1996). In contrast, Ericsson, Krampe, and Tesch-Römer (1993) reported that the amount of practice predicted a successful musical career. Middle-aged violinists were asked to estimate the accumulated hours of practice since beginning training. By age eighteen, the most promising students had invested 7,410 hours of practicing, while good students reached 5,301 hours and music teachers 3,420 hours. Similar results were obtained for pianists. Moreover, children who start earlier have more chances to reach a high level: whereas amateur musicians begin their musical training around age ten on average, professional musicians start around age six. Not any type of practice will do either: Ericsson and collaborators (1993, 1996) suggested that deliberate practice—a type of planned, strategically targeted practice, with clear goals such as concerts, competitions, and examinations—is much more conducive to future success than nontargeted practice.

However, Ericsson’s view has been challenged by other studies. Thus, scores on musical aptitude tests are also correlated with general intelligence (Lynn, Wilson, & Gault, 1989), even when socioeconomic status is held constant (Doxey & Wright, 1990). More recently, Meinz and Hambrick (2010) reported that although deliberate practice accounted for approximately 50 percent of the variance in piano sight-reading skill, working memory capacity, which is highly heritable and stable over time, as well as uncorrelated with deliberate practice, was also a significant predictor.

As we have seen, genes and heredity appear to play an important role in the development of musical ability, but environmental influences also exert a non-negligible influence, not only directly but also indirectly by modulating genetic influences. Recently, studies have explored the interaction between genes and environment by evaluating the heritability of behaviors such as musical practice in relation to objectively assessed musical aptitude. In a sample of 10,975 Swedish twins described in the previous section, Mosing, Madison, Pedersen, Kuja-Halkoka, and Ullen (2014) found that the amount of musical practice was substantially heritable (40 to 70 percent), and that the associations between musical practice...
and musical aptitude (as determined by SMDT scores) were mostly hereditary. Importantly, the results indicated that, when controlling for genetic influences, the amount of practice was not associated with musical aptitude. A follow-up study showed that shared genetic influences are responsible for the association between openness, the amount of musical practice, and the proneness to music-induced flow (Butkovic, Ullen, & Mosing, 2015). Another study by Hambrick & Tucker-Drob (2015) investigated the heritability of musical practice and musical achievement in a sample of 850 twin pairs and reported that genetic influences on musical achievement were strongest among individuals who engage in music practice, thus suggesting the presence of a gene-environment interaction.

There seems to be an optimal time window for music learning. Indeed, Gordon (1984) reported that scores on the Intermediate Measures of Music Audiation (IMMA) tend to stabilize around eight to ten years of age and show little improvement thereafter, a finding that was replicated using a different test battery (KMT) by Karma (2007) and Pulli et al. (2008). Furthermore, brain changes in the corpus callosum are observed only in musicians who started before age seven (Schlaug et al., 1995). Similarly, Elbert, Panter, Wienbruch, Rockstroh, and Taub (1995) found that only string players who began training before age seven showed enlarged brain areas involved in the representation of the little finger and the thumb of the left hand.

Only a small percentage of the general population shows extremely high musical abilities (Persson, 2009; Oerter, 2003). Early markers of highly talented children include a high sensitivity for musical structures (tonality, harmony, rhythm) and very good perception of musical expressions, whereas motoric aspects appear to be less crucial at a young age (Winner & Martino, 2000). Haroutounian (2000, 2002) points to perceptual awareness discrimination skills, as well as creativity, as other potential markers. Winner & Martino (2000) report that already at the age of two, talented children can imitate pitches or melodies after hearing them only once. Gagné (2003) emphasizes personal characteristics, such as motivation, temperament, time-management, physical characteristics, such as handspan, and the influence of the environment, including culture, parents, teachers, and peers. Although music competitions have traditionally been used to detect highly talented musicians at a young age, the advent of inexpensive genetic testing has led Levitin (2012) to propose identifying budding prodigies by their genetic pedigree.

While longstanding debates regarding the respective contribution of “nature” (innate talent, hereditary predispositions) versus “nurture” (training, environmental, cultural, and social influences) in the development of musical talent are by no means resolved, it is fair to say that the recent developments in both large-scale genotyping and large-scale assessment of musical aptitude promise to revolutionize the field. Already, a handful of pioneering studies published over the last few years suggest that the role of gene-environment interactions is much more complex than previously thought, and thus a reevaluation of seemingly established theoretical
concepts associated with musical expertise may be required. Perhaps the clearest illustration of this state of affairs is the finding that the ability or motivation to practice an instrument for long hours (which until recently would have been unambiguously classified as belonging to the “nurture” category) is itself subject to strong hereditary influences (Butkovic et al., 2015; Hambrick & Tucker-Drob, 2015; Mosing et al., 2014)—a finding that indicates that the “nature/nurture” distinction may in fact be at best misleading, if not entirely misconstrued.

**General Discussion and Conclusions**

In this section, we provide a critical appraisal of current views and measures of musical ability and offer some suggestions for future research. Although much remains to be examined, musical ability research has made considerable progress over the past decade. For instance, there is now evidence substantiating the very notion of aptitude, that is, the idea that certain musical skills are heritable. To anyone acquainted with child prodigies from Mozart to Michael Jackson, such a finding might seem obvious. Yet, recent studies go beyond confirming a role for heredity: they offer insights into gene-environment interactions and provide preliminary findings regarding specific alleles involved in the hereditary transmission of musical ability. Research on the genetics of musicality has also contributed to particularize the role of practice and training in the development of musical achievement. The studies in question are also relevant for a better understanding of the benefits that training of musical abilities may have for the enhancement of other, nonmusical skills (see Schellenberg, 2016b).

As noted at the outset, the credibility or interpretation of these findings is contingent on the methods for assessing musical ability. The recent advances in the development of tools for assessing musical ability are promising, but they also continue to highlight the challenges involved in creating sound diagnostic procedures and instruments. The two perhaps most important ones relate to sensitivity and validity. In any binary classification method, sensitivity and its counterpart, specificity, are key considerations. Unfortunately, it remains unclear how well classification procedures that dichotomize samples into musicians and nonmusicians are doing with regard to these indicators.

One reason for concern is the rate of “false negatives” produced by persons without musical training who nevertheless show above average performance on relatively comprehensive tests of musical ability (Kunert et al., 2016; Law & Zentner, 2012; Zentner & Strauß, 2017). Consistent with evidence from behavioral and molecular genetics, these “nonmusicians” are likely endowed with high, if undeveloped musical aptitude. The musically talented nonmusicians will likely boost the performance of the nonmusician group on any outcome that is actually affected by musical ability (e.g., certain language skills). Group differences between musicians and nonmusicians will thereby be reduced, resulting in underestimates of effects.
attributable to musical ability. By measuring musical skills directly, objective aptitude tests can help avert such interpretative biases.

Standardized aptitude tests should also help the interpretation of results related to effects of musical training. For example, a recent study suggests that associations between music training and intelligence can be largely accounted for by individual differences in music aptitude (Swaminathan, Schellenberg, & Khalil, 2017). The finding underscores the importance of controlling for initial levels of musical aptitude when studying effects of musical training on cognitive or other performance outcomes. Another difficulty in interpreting correlations between music training and nonmusical abilities is that children with high levels of music aptitude might be more likely than other children to take instrumental music lessons (Schellenberg, 2016b). If so, benefits of taking music lessons might be attributable to pre-existing levels of musical aptitude rather than to any effects of the music lessons. Another interpretational challenge in studies on the effects of musical training is whether benefits of music training are due to improvements in musical ability or to other, nonmusical effects of the musical training. Disambiguating the source of the effects can be achieved by measuring improvements in musical ability during the course of the study, which is possible via administration of musical aptitude tests at the beginning and at the end of the study.

As we have seen, researchers now have an array of musical aptitude tests at their disposal to strengthen the design of their studies and enhance the interpretability of their findings. Even so, the tests are not without their limitations. A particular concern with current batteries relates to limitations in content validity, that is, the extent to which they represent all relevant facets of musical ability. The range of musical skills currently captured by objective tests tends to be restricted to a small number of perceptual skills, usually acuity in the perception of certain rhythmic and tonal sequences. The PROMS assesses perceptual skills across a broader, but still incomplete range of musical domains. As long as musical ability is defined as performance on a few discrimination tasks, a conclusion such as “no causal effect of music practice on music ability” (Mosing et al., 2014) may or may not generalize to real-life expressions of musical ability. Researchers should remain cognizant of these limitations and be specific about the subset of musical skills to which their findings may actually apply.

To some extent, these concerns could be alleviated by evidence supporting the tests’ criterion validity. Intelligence tests offer an interesting point for comparison. It is well known that IQ tests do not measure all intelligent behaviors. The Raven Progressive Matrices test, for example, consists of abstract visual patterns that do not stand out as obvious items for the measurement of general mental ability. Yet, that is precisely what the Raven test has been shown to assess and predict in hundreds of studies (see Jensen, 1998, for a review). Similarly, single-letter knowledge and phoneme discrimination are among the most sensitive predictors of broader measures of linguistic proficiency, such as reading ability (e.g., Kirby, Parrila, & Pfeiffer, 2003; Muter, Hulme, Snowling, & Stevenson, 2004).
Understanding just how well the rudimentary discriminations required in current musical aptitude measures predict musical proficiency more broadly would mark a significant step forward. A study cited earlier provides an example of how research along these lines could progress. Kunert et al. (2016) examined criterion validity of the Brief-PROMS against performance on a harmonic closure task, and found associations between performances on both measures to be moderately strong (see table 16.1). Because the PROMS does not directly measure harmonic understanding, the result indicates that the PROMS may capture some basic aspect of musical ability that generalizes to other, more complex ones. More studies along these lines are necessary to gauge the extent to which current musical aptitude measures are reflective and predictive of broader musical capacities.

Three types of additions should be particularly helpful in bolstering the validity of current measures of musical ability: measures of the capacity for musical enjoyment, tasks that tap into musical production skills, and tasks that provide insight into musical creativity. As discussed in the section on “The Concept of Musical Ability,” a few attempts to develop the former two measures have been undertaken, but further work is necessary to make them fit for inclusion in general assessment batteries. Measuring musical creativity in musically untrained individuals represents a greater challenge. The use of intuitive touchscreen interfaces that allow participants to create musical sounds or to organize musical patterns into larger wholes could offer interesting possibilities. Yet, participants’ “creations” would still have to be evaluated against some standard of musical creativity, for which a consensus might not be easily achievable.

Beyond gains in validity, the inclusion of additional test components will allow researchers to analyze relationships among skills in music perception, music production, and the capacity for musical enjoyment. Through the expansion of current test batteries we will be able to understand how skills in rhythm perception are related to skills in rhythm reproduction, such as measured via beat tapping tasks. We will also be in a position to gauge the extent to which the capacity to enjoy music is related to objectively assessed musical capacities. In turn, progress in assessment tools will allow researchers to examine more complex research questions, including questions about the genetics of various musical skills or their relationship to other skills. As will have transpired in this chapter, research on the assessment of musical ability extends well beyond psychometrics, providing a window into the nature of musicality itself.

Note

1. The updated DTT can be found online at https://www.nidcd.nih.gov/tunestest/take-distorted-tunes-test.
References


The Assessment and Genesis of Musical Ability and Its Determinants


