Creating tests of musical ability for the 21st century

DANIEL MÜLLENSIEFEN

1. Introduction

In his very productive work life Heiner Gembris has written extensively (e.g. Gembris, 1997; Gembris, 1998) about individual differences in musical abilities and musical development as well as about tests that aim to measure talent and giftedness in music. In his critical assessment of established tests batteries Gembris has pointed to several shortcomings of tests that are affiliated with what he calls the ‘psychometric tradition’ (Gembris, 1997). These tests are characterised by procedures to score the performance on musical listening tasks in a deterministic way that is claimed to be objective. This includes the famous tests by Seashore (1919), Wing (1968), Bentley (1966), and Gordon (1965; 1979; 1982; 1989).

Gembris makes several critical points, also addressing the low validity and reliability of many established tests. But one of the main issues he raises concerns the nature of the tasks that are designed to assess a musical skill or ability. Many musical test batteries adhere to the design paradigm already promoted by Seashore which targets ‘atomatic’ musical skills using fairly abstract and generic stimuli. This stimulus design paradigm has the big practical advantage that stimuli are easy to create. But, more importantly, Seashore argues that the use of abstract auditory stimuli enables the measurement of ‘pure’ musical talent or potential, which is biased by any prior musical experiences or training that participants might have had.

However, Gembris critically argues that generic tone sequences or rhythms that are stripped of all but one musical parameter can hardly be considered valid musical stimuli representative of the rich musical world that children and adults in Western cultures are exposed to and learn from. Therefore, the constructs that these tests are measuring might not be musical skills but
rather abstract non-verbal auditory capacities. Hence, creating deliberately ‘unmusical’ stimuli for a music test to avoid the bias from prior musical experiences is throwing the baby out with the bathwater.

Gembris’ criticism is still valid today and also applies to many tests that have been published as part of the new surge in musical test development over the last decade (e.g. Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010; Law & Zentner, 2012; Ullén, Mosing, Holm, Eriksson, & Madison, 2014). These tests explicitly (Law & Zentner, 2012) or implicitly (Wallentin et al., 2010) aim to assess low-level ‘atonistic’ abilities that often target general auditory rather than specific musical abilities. In addition to the conceptual difficulties with this design paradigm, the use of generic and artificial auditory stimuli often leads to relatively unengaging and tedious test batteries which are especially difficult for children or participants with low concentration spans and motivation.

But if the creation of tests with artificial stimuli is not the answer to measuring ‘pure’ musical aptitude or potential (i.e. discounting the effect of musical training), then what is it?

One option we have suggested is not to discard the influence of musical training but, quite the opposite, to take musical training into account when making inferences from observable musical test performance to underlying musical potential. This was one of the main aims for creating the Goldsmiths Musical Sophistication Index (Gold-MSI, Millensiephen, Gingras, Musil, & Stewart, 2014), a test battery and corresponding self-report inventory that has grown into a new paradigm for constructing modern tests of musical abilities.

2. Developing the initial Goldsmiths Musical Sophistication Index (Gold-MSI)

The initial scientific motivation for designing a new self-report inventory on musical expertise, skilled behaviours, and musical training was born out of the frustration with the insufficiency of musical training questionnaires available up to 2010. The general practice in music research labs around the world mainly consisted in using informal ad-hoc questionnaires that had never been systematically constructed or validated. Examples of ad-hoc questionnaires that we obtained from labs in Canada, the US, and Europe typically contained a multitude of questions with different response options (e.g. rating scales, binary choices, multiple categorical choices, free-text boxes) and without any indication of how the data should be scored in a systematic way. From many publications it seemed that data for subsequent analyses was taken primarily from a question on the number of years of formal musical training. This variable was then often binarized to obtain a classification of participants as musicians and non-musicians. As Daly and Hall (2018) show the practice of binarizing the number of years of musical training is often done in arbitrary ways and can have detrimental effects for effect sizes and the significance of statistical models. The exclusive focus on formal training also ignores other types of musical learning besides formal music lessons. In addition to these informal questionnaires widely used in music research practice, we were aware of a few inventories with greater internal coherence, among them the Ollen Musical Sophistication Index (OMSI, Ollen, 2006). The OMSI was the result of a PhD thesis on the musical training experience of music students and hence its generalizability for populations without significant musical training is uncertain. The OMSI was never published in a peer-reviewed journal, but more importantly the suggested scoring scheme did not make intuitive sense, due to the underlying logistic regression model that seemed to overfit the data.

Hence, together with my colleagues at Goldsmiths I saw a need to construct a new self-report inventory on musical skills and expertise from scratch. Luckily, when the first work on the Gold-MSI self-report inventory had just begun, we were approached by the BBC’s citizen science unit, BBC Lab UK, who were interested in launching a new audience participation experiment. This provided an absolutely unique chance to construct and calibrate the new self-report inventory with a very large sample from the UK’s general...
population. To make this audience participation experiment engaging and fun for the users, BBC LabUK also suggested to create a small suite of ‘musical puzzles’ using real music clips that users would be able to take and get feedback for in an online implementation. For us, this was a chance to create new listening tests that had to be fun and engaging and that could also be linked to and validated parts of the self-report inventory. The audience participation experiment was named How Musical Are You? and in addition to the self-report questionnaire it comprised four tests, assessing melodic memory, beat perception and beat tapping ability, as well as sound similarity perception. The combination of a new self-report inventory and a suite of listening tests provided the opportunity to establish a clear link between subjective self-reports and objectively measurable listening abilities. Thus, combining these two types of data then opens up a new perspective for investigating interesting questions, e.g. on the impact of musical experience on different listening abilities or on the congruence of self-reported and objectively measured musical abilities.

The development of the Gold-MSI self-report inventory and its associations with listening test scores as well as with different kinds of demographic information is reported in Müllensiefen et al. (2014). While publishing the self-report inventory together with new listening tests was an important first step, we realised over the next few years that the initial battery of listening tests as implemented by BBC LabUK had some important shortcomings.

3. What is wrong with traditional and current musical listening tests?

The limitations and shortcomings that we noticed with the initial Gold-MSI listening tests are by no means unique to the Gold-MSI tests but apply almost equally to traditional test batteries from Seashore to Gordon as well as to recently published batteries. These shortcomings are not new and most of them are occasionally discussed in the literature (e.g. Karma, 2007; Hemmings, 2014). But for changing our perspective on musical ability testing, it was a crucial to get a first-hand experience of the limitations of the tests we had created ourselves. Hence, in the following I am going to describe limitations as they apply to the initial versions of the Gold-MSI listening tests but it is helpful to know that the same limitations apply to almost any other musical listening test battery as well.

3.1 Fixed test length and limited discriminatory power

A severe constraint of the tests was their fixed length and use of only a small set of items (e.g. musical stimuli). A fixed test length means that a test cannot be shortened if time in a test session is scarce and it cannot be lengthened to measure a specific ability with greater precision. If a test has only 18 items like the initial version of the Beat Perception Test and the response format is a 2-alternative forced choice task (2AFC, giving a 50% chance of guessing the correct response for each item), then there is only a limited range of meaningful scores, ranging from guessing level at 9 to perfect discrimination at 18 score points. Thus, even with perfect measurement precision, a test with 18 items using a 2AFC format can only discriminate between 9 different ability levels in any participant population. This discriminatory power might be sufficient for a coarse look at relationships within the general population, but it is insufficient for studying performance differences at the high or low end of the ability spectrum (e.g. among professional musicians or young children).

3.2 Moderate psychometric benchmarks and measurement error

The benchmarks for internal consistency, test-retest reliability and concurrent validity of the listening tests reported in Müllensiefen et al. (2014) were all moderate in size. This suggests that the tests measure with a considerable amount of error. Measurement error arises at least when tests are short (only have few items) and the chance of guessing responses of individual items correctly is high. Moderate psychometric benchmarks are common with other short tests using the 2AFC response format (e.g. PROMS test, Law et
Zentner, 2012), but less so with longer tests (e.g. the Musical Ear Test, MET, published by Wallentin et al., 2010, has 52 items for each subtest). But guessing is hard to avoid for participants at the low end of the ability spectrum because most test items (calibrated for average ability levels in the normal population) will be too hard and hence low-ability participants have no choice but to guess their answers. Having many items that are too difficult for a specific subgroup of participants essentially wastes participant time and contributes to measurement error. Similarly having many items that are too easy for discriminating among high-ability participants also wastes their time.

The consequence of measurement error are attenuated relationships of test scores with other variables of interest for example in correlation or regression models. Thus, if you are not able to measure a musical ability with sufficient precision (i.e. but only with considerable amount of measurement error), then all relationships with other variables will appear weak. Attenuation can be mitigated against using attenuation correction measures, but this is rarely done in practice, and if at all only for correlations (see e.g. Hambrick et al., 2014; Platz, Körner, Lehmann & Wolf, 2014). Accounting for measurement error in predictor variables as part of regression models is rare, even though it can be extremely important for safeguarding against wrong and biased conclusions (e.g. Jerrim & Vignoles, 2013; Goldstein & French, 2015). Hence, a precise quantification of the error associated with the measurements of a musical ability test is an important piece of information for constructing powerful models of musical behaviour.

### 3.3 Unclear construct validity

Often, the published documentation for many musical ability tests only contains a very vague description of the concrete abilities and constructs that the tests are supposed to measure. In most cases a cognitive process model, that would explain participants respond to a test item, is missing. For example, for the initial version of the melodic memory test of the Gold-MSI, it was obvious that items would require the processing of melodic information and the use of a memory component. But exactly how the processing takes places and whether there is only one possible cognitive strategy or several strategies to process the item correctly, is mainly unclear. But if it is unclear how participants process the stimuli of a test, then it is also unclear what the test actually measures (see Embretson, 1983, and Borsboom, Mellenbergh, & van Heerden, 2004, for similar discussions on test validity). An example is the initial version of the Gold-MSI melodic memory test that uses a same-different paradigm (see Müllensiefen & Hemming, 2018, for a description of the paradigm) with variants of the same short melody that need to be compared. The paradigm is essentially the same for melodic memory tests on almost any other test battery (Seashore, Gordon, PROMS, SMDT) and it assumes that participants commit a sequence of pitch intervals to memory. But if the two variants of the short melody are not transposed, a possible alternative strategy is to memorise the absolute pitch values of the first variant and compare them to the set of absolute pitches of the second melody. In this case the test does not measure the ability to memorise melodic structure (i.e. sequence of pitch intervals) but to remember absolute pitch values. This example shows that, unless a cognitive process model is described, it is impossible to know and to discuss what a given test is measuring. A cognitive process model would also typically describe features of the stimuli or the test procedure that make items easier or more difficult to process. These descriptions can be turned into empirical hypotheses and tested experimentally (Harrison, Musil, & Müllensiefen, 2016). This allows to test the assumptions that an ability test rests on and also helps to identify mechanisms that can be used for systematic construction of items covering a wide range of difficulties.

The first-hand experience of these limitations lead us to reconsider the process and the goals of test construction which enabled the re-design the initial Gold-MSI musical ability tests, overcoming many of the limitations listed above. The new approach we used in the redesign of the existing tests eventually gave rise to a new paradigm for musical test construction that we
developed over recent years and have been applying successfully for the construction of a series of new tests as well (Harrison, Collins, & Müllensiefen, 2017; Harrison & Müllensiefen, 2018; Larrouy-Maestri, Harrison, & Müllensiefen, 2019).

4. Developing the second generation of Gold-MSI musical ability tests

With the experience from the initial collection of Gold-MSI tests we were able to formulate very clear goals in terms of the features that the new musical ability tests should possess.

First of all, we aimed for short tests that provide precise measurements and hence are maximally efficient (i.e., they do not waste participant’s time). Considering the natural trade-off between test length and measurement precision the overall goal was to create flexible tests that could be shortened to provide a coarse (i.e., less precise) snapshot measurement or lengthened to measure with a precision necessary for in-depth studies of a specific ability. The prerequisite for this flexibility is obtaining a precise quantification of the test’s measurement error at each possible test length (i.e., number of items).

Secondly, for each test we aimed at constructing an underlying cognitive process model that makes the cognitive assumptions of the test implicit and contributes to its construct validity.

Furthermore, we were mainly interested in assessing higher-level musical abilities that would be complementary to the atomistic approaches of low-level and rather abstract musical abilities already out in the published literature. Since we had noticed that part of the attractiveness of the tests used with BBC LabUK was their use of a stylistically diverse range of audio clips, we also strove to make use of real rather than artificially created music where possible.

Finally, we realised that for high precision measurement at all possible ability levels and for repeated tests of the same participants it is necessary to have a large pool of items to choose from. But because creating a large item pool by hand is tedious and error-prone we opted for automated item construction using insights from the cognitive process model and empirical evidence of musical features that affect item difficulty.

Luckily, modern item response theory (IRT) provides a methodological framework that allowed us to work with approaches of cognitive process models, automated item generation and computerised adaptive testing in an integrated way. Item response theory originated in the 1960s (Rasch, 1960, see overview in de Ayala, 2009) and describes the relationship between item difficulties and participant abilities in probabilistic models. IRT models allow to estimate the ability of a participant after each trial of a test which then provides the opportunity to choose an item with a difficulty that matches the estimated participant ability as closely as possible. Hence the information gained from the response to an item is maximally informative and avoids the presentation of items that are either too easy or too difficult for a participant. This is the basic mechanism of the computerised adaptive testing (CAT) procedure which ensures that a test is highly efficient and that the length of the test can be chosen flexibly. Also, test measurements from IRT models are always associated with the corresponding estimate of their measurement error. Details of the IRT framework that we use and of its core ingredients (cognitive process modelling, automated item generation and computerised adaptive testing) are further discussed in Harrison, Collins & Müllensiefen (2017) and Harrison & Müllensiefen (2018). But one of the most appealing aspects of the framework is that it provides a synthesis of the two main traditions in psychological research, cognitive/experimental psychology and individual differences psychology which will be explored in the next section.
5. How combining the two big traditions of psychological research can inform a new approach for creating musical ability tests

Several authors have described and lamented the schism of the two big traditions in psychological research in the 20th century, namely experimental/ cognitive psychology and individual differences psychology (Boorsboom, 2006). But in addition to the fundamental problems for psychology as a discipline, music psychology provides a very good example of why this schism and the mutual ignorance of the work from these two camps is detrimental to new development in musical ability testing.

In a nutshell, cognitive and experimental psychology aim to answer the question why different stimuli elicit different responses. This question rests on the assumption that different people give similar responses to the same stimuli and that differences in their responses are largely noise. In contrast, individual differences psychology asks the question why different people give different answers to the same stimuli. Thus, what is considered noise in experimental psychology represents meaningful data to individual differences psychologists. The differences between the two camps are also reflected in the choice of their main analysis methods. While experimental psychologists look for differences between experimental conditions often using t-tests and ANOVAs, individual differences psychologists mainly rely on correlational methods to investigate the strengths of the similarities between variables. These differences in research methodology and differences regarding the nature of fundamental research questions has led to a situation where researchers from each camp work largely without reference to developments in the other camp. One example is the research on melodic memory that has a long and rich tradition in experimental music psychology (e.g. Dowling & Fujitani, 1971; Dowling, 1978; Ativa-Kabiri, Vecchi, Granot, Basso, & Schöö, 2009; Schulze et al., 2012) and has informed us about impact of different musical features (e.g. contour, tonality, length, complexity) on the ability to perceive, remember and retrieve melodies from memory. On the other hand, melodic memory tests have been a part of almost all musical test batteries published since 1919 and are an integral part of the individual differences research tradition within music psychology. They aim to assess melodic memory performance ability of an individual or the potential for future musical achievements. While cognitive experiments such as the ones by Dowling and colleagues (Dowling & Fujitani, 1971; Dowling, 1978) and the melodic memory tests on the tests published by Seashore, Bentley, Gordon, and Law and Zentner use exactly the same test procedure and generate the same kind of data, publications from the two camps hardly ever reference each other or make productive use of the insights generated in the other camp.

However, we found that insights from both camps can be used productively for the creation of new musical ability tests. Cognitive process models can be constructed around cognitive models from experimental psychology, for example by incorporating the features of musical structure that have proven to affect melodic recognition performance most strongly. In addition, the feature-based analysis of melodic items from existing melodic memory tests is a first step towards the calibration of a new set of items for which item difficulty is already known. A key tool for bridging the knowledge from both camps are explanatory item response models (deBoeck & Wilson, 2004). Technically these models are general linear models for binary responses and as such have been part of various statistical toolboxes for a long time. But the understanding that these models can explain the difficulty of individual items in terms of their (musical) features is rather recent. This function of general linear models is extremely useful for calibrating a test and also for predicting the difficulty of new items that have never been evaluated with a sample of participants. Thus, a highly predictive and reliable explanatory item response model (i.e. general linear model for binary responses) is the key for creating a large item pool and for scoring participant performance in computerized adaptive tests. Thus, explanatory item response models establish a clear relationship between the empirical knowledge on music perception and cognition and its application for the
precise measurement in individual differences research. As we have shown (Harrison & Müllensiefen, 2018; Harrison et al., 2017), these models can then confirm and expand knowledge from music cognition and perception with data gathered from individual differences tests.

6. So what?

Applying this new paradigm for constructing musical ability tests has enabled us to create a range of new tests over recent years, each targeting a different musical skill. This includes a melodic discrimination test (Harrison et al., 2017), a beat perception test (Harrison & Müllensiefen, 2018), a mistuning perception test (Larroy-Maestri et al., 2019), a rhythm ability test (Müllensiefen, Fiedler, Andrade, Forth, & Fricke, 2018), pitch imagery test (Gelding et al., in review), and a test for musical emotion discrimination (MacGregor & Müllensiefen, 2019). Further tests covering still other skills and abilities are currently in preparation. All of the Gold-MSI test are released as free tools and can be run within R, either as stand-alone implementations for testing at a single computer or as server-implementations for online testing.

But now that efficient and reliable tests of higher level musical abilities are available, what are the big questions that we can hope to answer?

The list of research questions where tests of musical abilities can play an important role is potentially very long. But some questions seem to suggest itself as primary because they can help to understand the scope and nature of the abilities that can be measured using the new generation of Gold-MSI tests.

A first question concerns the structure of musical abilities and how different musical abilities relate to each other. Over the course of the last century different answers have been suggested, ranging from a "set of largely unrelated abilities" (Seashore, 1947) to a strong primary component that connects all musical abilities Wing (1968), similar to the g factor in intelligence research. Larger factor analytic studies of auditory and musical tests (Stankov & Horn, 1980; Carroll, 1993) seem to point towards the importance of a central ability component. But the data in those studies was mainly obtained from auditory but non-musical tests or test batteries that follow that atomistic paradigm of abstract and generic musical stimuli (i.e. Seashore, 1919). But would a central musical ability component still explain most of the variance in a test battery that comprises tests of higher musical abilities ranging from pitch imagery, rhythm memory and emotion discrimination? This is an empirical question that we can start to address now.

A second question concerns the development of musical abilities, an area that has been a core aspect of Heiner Gembris' work (Gembris, 1998). Unlike for other psychological factors such as intelligence, working memory or personality, there is no quantitative longitudinal data available that would describe the development of individual musical abilities or for musicality in general. The lack of this kind of longitudinal data was one of the main motivations for starting the LongGold project, a longitudinal project on the development of musical abilities across adolescence. The LongGold project started in 2015 in the UK (Müllensiefen, Harrison, Caprini, & Fancourt, 2015) and has a parallel branch in German since 2018. It assesses secondary school children once a year, testing their musical abilities but also taking measures of intelligence, personality, attitudes as well as socio-emotive measures. The same children are assessed every year and hence the linking of data across time then allows for the construction of developmental curves. The test battery comprises 20 measures and the annual test slots are only 90 minutes long. This means that each of the 20 tests has to be maximally efficient and hence the computerised adaptive version of the Gold-MSI tests are an ideal toolbox for this research scenario. A very concrete outcome of this project will be the construction of age- and training-dependent benchmark data for different musical skills. This benchmark data can then serve for comparisons in future studies assessing the same or related musical skills or abilities.
Finally, the use of the Gold-MSI tests along with tests of other cognitive abilities within the LongGold project also offers the possibility to answer interesting questions on transfer effects from music to other skills (i.e. intelligence → ‘Does music make kids smarter?’) as well as transfer effects from other kids to musical abilities (i.e. → ‘Do smarter kids make more music?’). These questions on the direction of causal effects have been hotly debated for a long time in the music psychology and music education communities but are very difficult to answer with certainty from cross-sectional data or data from specific short-term interventions. However, the perspective changes substantially if longitudinal data is available because then it is possible to observe temporal patterns, e.g. potential increases of musical activity during adolescence that might be followed by subsequent change in intelligence, personality, and attitudes.

While it will take time to arrive at answers to these questions, it is important to remember that basis for these research endeavours is the existence of valid, reliable, and efficient tests of relevant musical abilities. Hence, while the psychometrics of music does not immediately strike as the sexiest sub-discipline in music psychology, it provides the basis for sound empirical data and is therefore a crucial part of the research process that delivers evidence for or against theories on the nature of human musicality.

References


---

**What needs to be concluded from the fact that there is no quantitative measure for musicality?**

**JAN HEMMING**

To Heiner Gembris’ 65th birthday and the 100th birthday of measuring musicality.

John Sloboda, Jane Davidson and Michael Howe once claimed “Self-beliefs in one’s musicality are better predictors of high performance in music than tests of musicality are” (Sloboda, Davidson, & Howe, 1994, p. 364). While they are likely to be right, it is telling that they do not give any reference, albeit empirical evidence for this statement. But how could they have done so? And would we be able to empirically support this assumption nowadays? 1994 was around the time when the musicality debate started to rise again in the light of the expertise-paradigm, culminating in the notion "Musicians: Experts not geniuses" which is the title of the paper just cited.

In the years before, music psychology and pedagogy had largely tried to keep some distance from notions like giftedness, talent and musicality. Above all, with the exception of the continuous work from Gordon (1965, 1979, 1986, 1993), researchers stayed away from testing procedures aiming at providing quantitative measures for musicality and the like. This is both true for the application of existing tests such as Seashore (1919), Wing (1948,1968), Bentley (1966) and others, as well as for the development of new tests. The rediscovery of musicality research had instead been prepared by qualitative approaches, mostly among outstanding achievers, internationally (e.g. Howe, 1996) as well as in the German speaking countries (e.g. Bastian, 1989). There is no doubt that individuals differ in their degree of musicality, just as they differ in their degree of mental or cognitive abilities.

---

1 While the terms musicality, giftedness and talent are not identical, they are used interchangeably in this text (see Hemming, 2004 for a detailed terminological discussion).