reflection, and rules that have a general rule in common combined with each other. When it comes to rotations, it is strongly recommended to work with two numerical values that are not so distant from one another (see the examples inside the package).

2. It is useless to create an object with the 'build_fa' function if this object is not then plotted. The easiest way to plot is by simply using the 'plot_fa' function. It is not recommended to save the plot in a directory by means of the R interface buttons; instead, a directory can be given as another argument so as to let the function perform this procedure precisely and straightforward.

Assessing melodic discrimination abilities with computerised adaptive testing and automatic item generation

by Peter Harrison, Tom Collins and Daniel Müllensiefen

The ability to process and understand music is a universal human faculty. However, individuals can vary immensely in their ability to process musical materials, and many tests have been developed over the past century to assess these differences. Historically, these tests have typically been used to measure musical 'aptitude', with the aim of selecting the most able children for music education and instrument tuition. However, musical ability tests are increasingly being used in psychological and neuroscientific research, investigating how musical abilities affect music cognition, and how individual differences in musical abilities relate to individual differences in other cognitive abilities.

At the **Music, Mind & Brain group** at Goldsmiths, University of London, we have aimed to develop a musical test battery and corresponding self-report instrument called the **Goldsmiths Musical Sophistication Index** (Gold-MSI). This instrument is intended not to rely on formal musical training, not to be biased towards a particular musical style, and to be representative of a large range of active musical behaviours in the general population.

It can be difficult to test for musical listening abilities effectively. If the test is to measure a broad range of abilities, as is common in the general population, it must contain items spanning a wide range of difficulty levels. This usually requires the test to be relatively long. Moreover, if listening tests work by aural presentation (and do not use musical notation), then they tend to provide only a small number of response options for each question. This contributes noise to test scores, because it is easy to guess the correct answer by chance. In order to compensate for this noise, test length needs to be increased further. However, musical listening tests are already tiring for test-takers, because test items usually take a long time to administer and demand a lot of concentration. This makes it difficult to increase test length further without increasing fatigue and hence diminishing reliability.

One possible way of addressing this problem is through computerised adaptive testing. Computerised adaptive tests (CATs) continuously tailor their difficulty to the estimated ability level of the test-taker as the test progresses (e.g. Figure 1). This can greatly increase test efficiency, as participants no longer have to take items that are far too easy or too difficult for their ability level. As a result, CATs typically can achieve the same reliability as traditional fixed-length tests even when test length is reduced by 50-70%.

Unfortunately, however, CATs can be very expensive to construct. CATs are usually built within Item Response Theory (IRT), a powerful psychometric framework for modelling ability tests. Each test item is modelled in terms of several psychometric parameters, such as difficulty, discrimination, and chance success rate. These parameters need to be estimated before the test can be used, but this can be a very expensive process, especially in the case of CATs, which typically require very large item banks.

One recently developed technique for making CATs more efficient to construct is automatic item generation (AIG). In AIG, item parameters are not estimated for each item individually, but are instead predicted on the basis of structural characteristics of the items. This relies on a good understanding of the cognitive processes behind test-taking and a clear conceptualisation of how individual differences in abilities can contribute to task performance. AIG can make CAT construction more efficient, because much less response data is required to calibrate an AIG model than would be required to estimate every item parameter individually.

We have applied these techniques to the construction of a test of melodic discrimination abilities. In this test, test-takers have to discriminate between three similar versions of the same melody, each of which is transposed slightly higher in pitch. In every

Music Testing and Bias

Many traditional measures of musical abilities rely heavily on skills that are typically acquired through formal musical training. This is problematic because formal musical training tends to be biased towards Western art music and typically relies heavily on knowledge of Western musical notation. While musical notation is certainly important for some musical styles and some modes of musical engagement, it is by no means important for all of them. For example, it is very possible to be an effective DJ, music journalist, or music producer without being able to read music.

See the Gold-MSI in Action

The test is implemented on the Concerto platform and an example implementation is available at http://concerto.icar-project.com/v4/?wid=5&tid=1.

trial, two of the melodies possess exactly the same interval content (i.e. the pitch relationships between notes are unchanged), whereas one melody has different interval content. The test-taker's task is to identify the 'odd-one-out', which can either come first, second, or third (Figure 2).

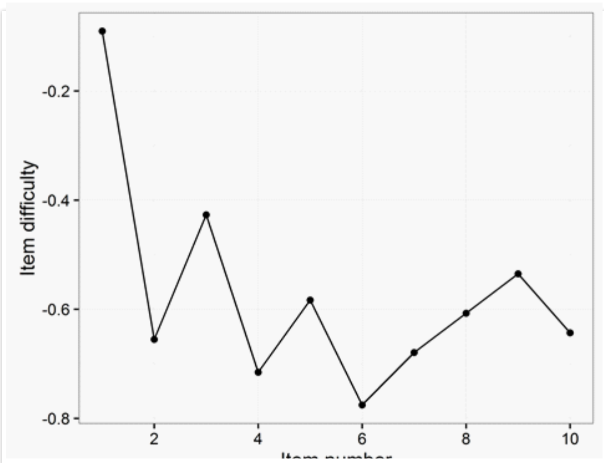


Figure 1
Simulation of the CAT item selection process for a low-ability test-taker on the melodic discrimination test. Over the course of the test, the CAT algorithm homes in on the test-taker's true ability level of -0.6.

We identified **four key cognitive processes** that underlie this test: perceptual encoding, memory retention, similarity comparison, and decision-making. In perceptual encoding, cognitive representations of the melodies are derived from the audio signal. These melody representations are retained in working memory, to allow the melodies to be compared. The comparisons take the form of similarity judgements between pairs of melodies. Finally, a decision-making process combines information from the similarity judgements to determine the final response.

This cognitive model forms the basis of our AIG system. Melodies for test items are automatically generated using an automatic composition algorithm, and the difficulty of these items is then predicted on the basis of our cognitive understanding of the task. Memory encoding difficulty is varied by manipulating the length of the melodies; longer melodies place higher demands on working memory, and so are harder to retain. Difficulty of similarity comparison is varied by manipulating the degree and type of differences between the melodies; more similar melodies are harder to discriminate.

We calibrated our AIG system using response data from 425 participants, each of whom took a 10-minute online melodic discrimination test. The AIG system was then used to construct a CAT with an item bank of 1,200 items. We then investigated how the test-retest reliability (Figure 3) and standard error of the ability estimates (Figure 4) of this CAT varied for different test lengths,

Figure 2

Example of one trial of the melodic discrimination test. Here the third melody is the 'odd-one-out'.

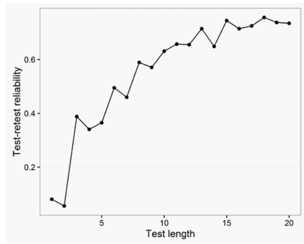


Figure 3

Test-retest reliability as a function of test length for the melodic discrimination CAT.

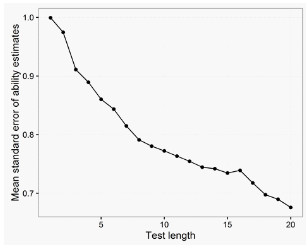


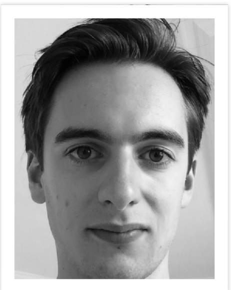
Figure 4

Standard error of ability estimates as a function of test length for the melodic discrimination CAT.

using a nationally representative sample group of 42 online test-takers. As Figure 3 demonstrates, a peak test-retest reliability of 0.75 is reached with about 15 items. On average, pre-existing melodic discrimination tests reach a similar reliability with about 30 items. Our approach therefore allows us to reduce test length by approximately half without compromising reliability. Test reliability should still higher under controlled laboratory conditions.

In the future, we hope to improve the melodic discrimination test further by extending the range of item difficulties available and improving item difficulty predictions. We are also developing a complementary beat perception test using similar techniques of computerised adaptive testing and automatic item generation. So far, our results suggest that these psychometric techniques have exciting potential for the future of musical ability testing.

about the authors



Peter Harrison

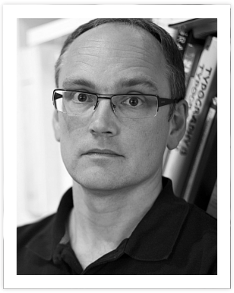
Peter Harrison is a PhD student at Queen Mary, University of London.

Contact Author via Email:
pharr011@gold.ac.uk

Tom Collins

Tom Collins is a Visiting Assistant Professor in the Department of Psychology, Lehigh University, Pennsylvania, USA. His research interests include the development of Web-based music software and its effect on student education and work, pattern discovery in music and other domains, modelling musical expectancy, and automatic identification of high-level music-theoretic concepts.

Author's website:
http://www.tomcollinsresearch.net/



Daniel Müllensiefen

Daniel is a music psychologist and member of the Music, Mind and Brain research group at Goldsmiths. His research interests include musicality and individual differences in musical abilities, the psychometrics of music, memory for music and melodies, involuntary musical imagery (i.e. 'earworms'), the perception of musical similarity, statistical models of cognition, music in advertising, corpus-based musicology, cognitive biases in musical judgement and cognitive issues related to music copyright.

Author's website:
http://www.doc.gold.ac.uk/~mas03dm/