

Development of a standard test of musical ability for participants in auditory interface testing

Alistair D N Edwards¹

Ben P Challis²

John C K Hankinson³

Fiona L Pirie

Department of Computer Science

University of York

Heslington

York

England

YO10 5DD

ABSTRACT

There is a danger that auditory interfaces will only be usable by people with musical skills. Researchers in testing new interfaces often try to control for musical abilities by classifying participants as ‘musicians’ and ‘non-musicians’. However, there is no agreement as to what constitutes a musician. It is proposed, therefore, that there should be a standard pre-test that can be applied to all participants. This paper describes a version of such a test which is being developed.

Keywords

Musical ability, evaluation, testing, musician, auditory interface.

INTRODUCTION

One concern in the development of auditory interfaces is that they will be usable only by people with particular auditory capabilities. In particular, there is the possibility that only people with musical skills will be able to use them. Many researchers have tried to take account of this in testing designs and ideas by classifying their test subjects as either ‘musicians’ or ‘non-musicians’ and then looking for differences in the results of the two groups [1, 2, 3]. There are a number of problems with these attempts, though. Firstly, there are disagreements as to how to define a musician. Secondly, most such definitions tend to be biased towards those with formal musical training and do not account for people who may have good but natural and untrained musical abilities. Finally, the results based on these definitions are equivocal. For instance, both Stevens (*op. cit.*) and Brewster (*op. cit.*) were concerned with how well people could make use of earcons, yet Stevens found that musicianship was *not* significant, whereas Brewster found that it *was*.

The basis of the work to be reported was that such arbitrary categorizations could be obviated by the development of a standard test of musical ability. We have dubbed the test the Musical Aptitude Test or *MAT*. This could be administered to all participants in auditory experiments, thereby providing a standard benchmark. Thereafter the results of individual participants in testing of auditory interactions could be related back to their *Mat* scores. For instance, poor performance by an individual on an auditory interface which placed a heavy reliance on rhythm could be explained by the fact that their *Mat* shows that they have a lower than average sense of rhythm. More to the point, if testing suggested that a particular design was usable only by people with high *Mat* scores, that would be an indication that the design was a bad one.

One of the main objectives in designing the tests was to be able to measure equally well the abilities of trained musicians and those who have developed musical attributes naturally. It is relatively easy to identify people who have had a musical education, have passed music exams and play an instrument. Yet there is a grey area (that the authors cited above struggled with). For instance, is a singer a musician? Surely someone who sings in a choir must have similar musical abilities to a player in an orchestra? But what if they only sing in the bath? Similarly, there are people who have never played an instrument, but who nevertheless spend a lot of time listening to and appreciating music. What abilities do they have – and how can we measure them?

No comparable tests exist already. The nearest equivalent is [4], but this was not suitable for our purposes firstly because it relies on some musical training and secondly because it is intended for the testing of children. In the UK the Grade tests of the Associated Board of the Royal Schools of Music are probably the most widely recognized measure of musical ability and they concentrate on the basic components of musical perception, such as rhythm, pulse, pitch, harmony, melody and listening skills [5]. However, they rely heavily on performance skills (clapping, singing or playing) and technical knowledge of music

1 +44 (0)1904 432775, Alistair.Edwards@cs.york.ac.uk

2 +44 (0)1904 432765, Ben.Challis@cs.york.ac.uk

3 +44 (0)1904 432748, John.Hankinson@cs.york.ac.uk

(terminology, interval labeling, chord identification *etc.*) Other, on-line tests exist (such as [6, 7]) but these all either assume an ability to read music notation or knowledge of a (music) keyboard or both.

There is some debate about the nature of trained and untrained musical ability. An attractive proposition is that the two forms are related to hemispheric specialization in the brain. Thus, one suggestion is that trained musicians rely more on the verbal processing which takes place in the left hemisphere. Such a musician, for instance, may be able to categorize a fragment of music and thereby attach a verbal label to it, while an untrained musician might not have that vocabulary and react more on an emotional level. The problem with such hypotheses is that attempts to verify them experimentally have proved equivocal. Some studies support localization of musical processing in the left hemisphere, some site it in the right hemisphere and some suggest it is cross-lateral! ([8] lists the relevant studies). Such theories do not help us. Hence we must fall back on the suggestion that there are some things that people who are 'musical' will be able to do better than those who are not, regardless of how they acquired that status.

This paper describes the design and initial evaluation of such a set of tests, which can be administered by computer. (Details of the design are available in [9]). These measure a set of component musical abilities and result in a profile of scores indicating comparative abilities in each component.

TEST DESIGN

The tests are presented on-line¹ and mostly consist of matching tasks. That is to say that a target sound is heard (usually twice) and then the participant must choose the matching example from a set of samples – without being able to re-listen to the target. In most cases a standard design of options is used, whereby some of the incorrect options are 'spoilers' which bare some resemblance to the target. A 'Don't know' option is always available, and participants are encouraged to use it in preference to making guesses. An example test is shown in Figure 1. No feedback is given as to the correctness or otherwise of the choice (since this is a testing, not a teaching, procedure).

The tests cover:

- 1 Pitch – pitch awareness;
- 2 Rhythm – duration;
- 3 Rhythm – meter;
- 4 Harmony – chord awareness;
- 5 Harmony – chord structure;
- 6 Rhythm – structure;
- 7 Pitch and rhythm – melody discrimination;
- 8 Dynamics – dynamic awareness.

These components are not necessarily musically comprehensive, but rather reflect the skills that are likely to be important in the use of any auditory interface. In all there are around 200 tests. Tests within a category vary in complexity. For instance Test 1 involves tone rows. As the test proceeds, the number of tones is increased from 1 to 8.

INITIAL RESULTS

So far the tests have been administered to a group of 30 people [10] . Most of the test participants were graduate students from the University of York. They were studying a variety of subjects and came from a variety of ethnic backgrounds.

This was effectively a pilot of the testing procedure. As will be discussed further below, the main problem in devising the test is its very novelty. With nothing to compare it with, it is hard to validate. Thus, the results of this pilot test will not be presented as establishing any kind of norm, but rather as raw data that can be manipulated and analysed in order to find out what is the appropriate way to interpret the results. There are also a number of practical short-comings of the test design which have shown up and which can be eliminated from the final version.

¹ The test requires a Windows™ PC fitted with a Midi sound card. The configuration used in these tests was a Pentium™ PC, running Windows 95 with a SoundBlaster™ PC128 sound card. The software was written in Borland™ C++.

The image shows a screenshot of a computer-based music test interface. On the left, a grey panel contains the test details: 'HARMONY - CHORD AWARENESS TEST', 'Question No: 7', and instructions to click a 'Test chord' button. Below are five numbered buttons (1-5) and a radio button interface for selecting an answer. On the right, a musical staff shows a 'Test chord' (F major triad) and five 'Specimen chords' (F major triad, F major triad, F major triad, F major triad, F major triad).

Figure 1. Sample exercise from Test 4, Chord Awareness. The participant hears the chord by pressing the *Test chord* button. When it has been played twice, the specimen chords can be heard by pressing the numbered buttons, and the answer selected from the radio buttons

Results on the individual tests were combined as appropriate, yielding five measured abilities:

- 1 Pitch (Test 1),
- 2 Rhythm (Tests 2, 3 and 6),
- 3 Pitch+Rhythm (Test 7),
- 4 Harmony (Tests 4 and 5),
- 5 Dynamics (Test 8).

A score is calculated for each of these, which is expressed as a ratio of the mean and normalized such that an average performance scores 100. For instance, the mean score in Test 1 was 73, so a score of 90 would be converted in to a 'Pitch Quotient' of $100(90/73) = 123$. It should be noted that because these test scores are self-referential there can be no direct comparisons across tests. So, for instance, it would not be right to suggest that the pitch perception of someone with a Pitch Quotient of 123 was 'as good as' their rhythmic abilities because they happened to have a Rhythm Quotient of approximately 120, or that it was 'twice as good as' their harmonic ability because that scored 60. Indeed, the scores have an ordinal dimension [11], so that a Pitch Quotient of 120 is better than one of 60, but is in no sense 'twice as good'.

IQ is a single number which is meant to characterize that property that we call intelligence. It would possible to aggregate scores in all categories to calculated an overall Mat musical quotient, but it was felt that the component abilities that make up musicality are too diverse for the Mat to be very meaningful so instead a five-part profile was constructed for each participant¹. The five scores calculated were designated: Pitch (P), Rhythm (R), Pitch+Rhythm (PR), Harmony (H) and Dynamics (D).

All the participants who undertook the tests were also given an extensive questionnaire. This covered their background both musically and non-musically (ethnicity, education, *etc.*). The questionnaire information can be used to try to account for similarities (and differences) in profiles. The questionnaire has already revealed some interesting patterns. For instance, nearly all the participants stated that they listen to music, so that this question is not yielding much information. Similarly, the participants were invited to categorize themselves, and only one of them used the classification, 'Non-musician and not interested in music'. This suggests that – at least among the population we used – most people prefer to profess some level of interest in music.

¹ Some people question the validity of a single measure of intelligence, too, and prefer to describe peoples' multiple intelligences [12].

Initial Mat results show some interesting patterns. For instance, eight people had consistently below average profiles. Six of them had no or minimal musical training and only one had ever played an instrument. At the other end of the scale, most of the 12 people with above average profiles could be described as trained musicians (have had lessons, play instruments or sing etc.), while two of them were musically untrained.

There were two people who scored above average in all the tests. They both had a musical backgrounds, but of different forms. Both have played instruments, but whereas one stills played regularly, the other's playing had lapsed – but was still a regular singer. Both listened to music frequently and both spoke several languages.

At the other end of the scale, three people scored below average in all tests. They all described themselves as 'Non-musician but interested in music'. Although two of them had at some time had played an instrument, none of them had had lessons nor played currently. None had any knowledge of music theory (reading or sight reading music) nor could they improvise. All of them listened to music.

VALIDITY AND RELIABILITY

The reliability of the test is being explored by retesting some of the original participants, some months after their initial exposure to the tests. If the test is reliable, they can be expected to obtain similar scores.

The distribution of scores on most of the tests appears to be approximately normal (as adjudged by comparing means, modes and medians and measuring skew). This is reassuring as there seems no reason to expect that the skills being tested would not follow a normal distribution.

As stated earlier, there is no existing test designed to measure both taught and innate musical ability. For instance a correlation between Mat scores and Associated Board grades might be reassuring, but will not account for good results by untrained musicians who have never entered for the Associated Board. We already have examples of acknowledged musicians who scored well in several of the tests but below average on one or two of them. We would contend that this is a vindication of the tests, that it reflects the fact that even a trained musician will not be equally good at all aspects of music. Nevertheless he hope to obtain further validation. For instance, we will apply the Bentley test [6] to some of our participants and look for gross correlations. We also plan to apply the test to people for whom we would predict particular results and verify those predictions. For instance, a percussionist would be expected to score well on rhythm, but perhaps less well on pitch.

FURTHER WORK

There is a need to refine the design of the tests. As they stand they take rather a long time (an average of 57 minutes). Considering they are meant to be preliminary to substantial testing of some other auditory interface, it would be desirable for them to take much less time. Cutting down the numbers of tests and exercises without losing significant amounts of information must be explored. It seems that some of the tests may have been too easy – all the pilot participants got them right – and these would be obvious candidates for pruning. Conversely, some of the current tests seem to be too hard and probably need to be toned down. In particular, scores in the Pitch and Rhythm test (Test 7) were rather low, making it hard to draw conclusions from the scores.

A simple modification that is needed is to re-design the layout of the dialogues. As shown in Figure 1, the layout of the test buttons is not congruent with that of the answer radio buttons. This is a simple potential cause of slips that can be eliminated.

There is also a question as to the number of times the test target should be presented. As described earlier, the participant usually had to hear it exactly twice. However, there are arguments to suggest that the user should be allowed to control the number of presentations. In other words, the instruction might be, 'Listen to the sample as many times as necessary for you to feel that you remember it, and then listen to the numbered specimens.' Another flaw in the pilot test was that tests were presented in increasing order of difficulty. Clearly this encourages learning. For fairer testing, the order of difficulty should be randomized.

Once these flaws have been ironed out and an appropriate method devised to analyse and present the results, the test will have to be administered to a very large number of subjects. Only then will a reliable baseline be established. We hope for the collaboration and co-operation of the ICAD community in moving on to this stage of the development. Thereby we hope to attain the goal of having a standardized test as part of all auditory interface testing.

ACKNOWLEDGEMENTS

Authors BPC and JCKH were sponsored by studentships from the UK Engineering and Physical Sciences Research Council (awards 96309035 and 97307727 respectively). The Mat Tests were designed and programmed by BPC and JCKH. The pilot testing was carried out by FLP.

REFERENCES

1. Stevens, R., *Principles for the design of auditory interfaces to present complex information to blind computer users*. 1996, DPhil thesis, University of York, UK.

2. Brewster, S.A., *Providing a structured method for integrating non-speech audio into human-computer interfaces*. 1994, DPhil thesis, University of York.
3. Fitch, W.T. and G. Kramer, *Sonifying the Body Electric: Superiority of an auditory over a visual display in a complex, multivariate system*, in *Auditory display, sonification, audification and auditory interfaces: Proceedings of the First International Conference on Auditory Display*, G. Kramer, Editor. 1992, Santa Fe Institute, Addison-Wesley: Reading, Massachusetts. pp. 307-325.
4. Bentley, A., *Musical Ability in Children and its Measurement*. 1966, London: George G Harrap & Co Ltd.
5. *Syllabus of Examinations, 1998 and 1999*. 1997, Associated Board of the Royal Schools of Music.
6. *Ear Training Website*, <http://www.earpower.com>.
7. *Ear training with Earmaster*, (1999), <http://www.earmaster.com>.
8. Read, H., *Dissociable symbolic processing abilities in a case of global aphasia: Evidence for autonomous subsystems within music processing*. 2000, University of Sheffield (in preparation).
9. Hankinson, J.C.K., B.P. Challis, and A.D.N. Edwards, *MAT: A Tool For Measuring Musical Ability*. 1999, Technical report YCS 322, University of York, Department of Computer Science.
10. Pirie, F., *Implementation and evaluation of a proposed measure of musical ability for auditory interface testing*. 1999, MSc (IP) Project Report, University of York, Department of Computer Science:
11. Zhang, J., *A representational analysis of relational information displays*. *International Journal of Human-Computer Studies*, 1996. **45**: pp. 59-74
12. Gardner, H., *Frames of Mind: The Theory of Multiple Intelligences*. 1985, London: Paladin.