Goldsmiths Musical Sophistication Index (Gold-MSI) v1.0: Technical Report and Documentation Revision 0.3

Daniel Müllensiefen, Bruno Gingras, Lauren Stewart, Jason Jiří Musil

October 9, 2013

Document revision history

0.1	Dec 2011	Covered self-report measure.
0.2	Apr 2012	Covered v0.9 and v0.91 objective tests.
0.3	Aug 2012	Covered v1.0 objective and self-report measures.

Test version history

Version 0.9	January 2011	Used with BBC LabUK survey "How Musical Are You"?
Version 0.91	March 2011	Extended version used at Newcastle Science Festival and for extended laboratory testing.
Version 1.0	August 2011	New version of self-report measure and shorter, calibrated objective tests.

Contents

List of Figures

List of Tables

Acknowledgements

Music means many different things to many different individuals. We are grateful to the following individuals and groups, who have helped us to find meaning in our heaps of notes, data and words:

Andrea Halpern, John Iversen, and Aniruddh Patel, whose work we have borrowed from in designing some of our tests; Goldsmiths, University of London for the crucial early career grant (and fish'n'chips on Fridays); BBC LabUK and especially Mike Orwell and Joseph Coulson for the enormous data-set; Audio Network Plc and especially Nick Woollard for letting us use their fantastic audio library tracks; all at Qualtrics, for supporting and hosting our online surveys; Alan Pickering for generously advising on statistical methods when all hope is lost; Thenille Braun, Mandi Goldberg, Alex Handler, Andre Lira, and Katharina Bauer, who have moved mountains (questionnaires and otherwise) in their time as our interns; those students who have already used the Gold-MSI in their projects, especially Fatima Al Toma, Amit Avron, Budr El Nusairi, and Monika Ruszcynski for their additional help. Even virtuosos can fudge a note or miss a beat, so we humble scientists apologise especially to anyone who has been omitted from this list by accident.

1 Introduction

The Goldsmiths Musical Sophistication Index is a psychometric tool for the measurement of musical attitudes, behaviours, and skills. It comprises a self-report questionnaire as well as a suite of music-psychological tests measuring different musical skills. This technical report documents the items and materials used in the Goldsmiths Musical Sophistication Index (Gold-MSI) as well as the analyses and choices made in its development. It should enable researchers to understand the design of the Gold-MSI and will allow the recreation of its components from provided source materials.

For more information on the Gold-MSI as well as for all necessary materials see the project page at: `http://www.gold.ac.uk/music-mind-brain/gold-msi/`.

1.1 Note on task versions

Version 1.0 reflects both a process of test design (creating optimal short versions of each measure) and also developing understanding of those facets of musical sophistication which are being measured. Version v0.91 is our implementation of a set of tests functionally similar to those used in our collaboration with BBC LabUK (which we have identified as v0.9), albeit with an extended set of objective task stimuli (genre sorting, melodic memory, beat perception) designed to cover a broader range of difficulty levels[1]. These data were used to make optimal short tests with difficulty calibrated for the general Western listening population—these tests are version 1.0. Version 0.91 therefore represents a step in the research and development of the test battery, rather than a complete test, and data norms are not available for it. For this reason, we have made version 1.0 amd 0.9 materials available in the publicly accessible repository and v0.91 materials are available on request. Work on the self-report questionnaire preceded development of the objective ability tests, so data norms (see *Appendix A*) are given both for initial pilot data and also subsequent versions respectively. Norms are not yet available for version 1.0 of the objective tests.

1.2 Note on software and analyses

Scoring or usage instructions are given for each of the tests[2] in the corresponding sections of the manual. In general, these take the form of either specially prepared Microsoft Excel spreadsheets or of sections of easily understood code, written in the `R` language for statistical computing. `R` is free software and can be obtained for many platforms from `http://r-project.org`. We have also found Excel spreadsheets to be a convenient scoring method due to the ubiquity of both free and paid software which can open the spreadsheets and understand the macros used for scoring, as well as peoples' general familiarity with the format. A list of spreadsheet software can be found at the following URL: `http://en.wikipedia.org/wiki/List_of_spreadsheet_software`. In the latter case, although the scoring functions could easily be implemented in many other ways, we hope that researchers not already familiar with the powerful capabilities of `R` will at least be tempted to try it. We provide a general scoring function which reduces the task of scoring a given test to four lines of code (one to load a data file, one to load the function, one to run the scoring, and one to write an output file). Finally, because experimental participants have often requested feedback on their test scores, we have included some hints on how this may be given.

[1] Please note that the beat production task has, at the time of writing, been implemented only as the v0.9 version on the BBC website.

[2] Scoring instructions are not yet released for the beat production task.

2 The self-report questionnaire

The purpose of the self-report questionnaire as a psychometric instrument is a) to quantify the amount of musical engagement and behaviour of an individual in its many possible facets and b) to record the self-assessed level of various musical skills.

2.1 Development of the self-report questionnaire

2.1.1 Literature review and critique

We reviewed two areas of the literature that we considered relevant for the development of a new self-report inventory and test battery for musical sophistication: a) literature on measures of abilities and expertise outside music which showed potential for adaptation to the music domain and b) music-specific literature on the measurement of ability, achievement and attitudes.

In terms of the existing non-music literature on tests and test construction for measures of abilities and expertise we identified different approaches that had been applied to the measurement of a variety of cognitive and sensory abilities such as a criterion-referenced test battery for badminton (Yau, 1999), a wine knowledge test (Hughson & Boakes, 2002), the role of passive perceptual learning on the ability to identify wines (Hughson & Boakes, 2009), expertise in the categorisation and representation of problems in physics (Chi, Feltovich, & Glaser, 1981), and computer programming (Weiser & Shertz, 1983). Taken together, these studies indicate that experts generally use a different representational system than novices, as well as different criteria to categorise domain-specific problems. The studies also suggest that the ability to make fine distinctions and to use a complex system of categorisations—which allows for an increased repertoire of appropriate response behaviours—is a fundamental manifestation of expertise. We make use of this insight as a core component in our concept of musical sophistication.

With regard to the measurement of musical abilities, musical achievement, musical experience, and musical attitudes we reviewed classic and recent musicality tests, many of which were constructed for use in an educational context (Seashore, Lewis, & Saetveit, 1960; Wing, 1962; Bentley, 1966; Gordon, 1989; Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010; see summary discussion in Boyle & Radocy, 1987), or for the identification of special populations (Peretz, Champod, & Hyde, 2003). However, we found that none of these tests would serve our idea of a measure of musical sophistication appropriate to the general population of a western society, where we cannot assume any form of musical training for a large part of the population. For instance, many of the tests involve listening exercises, where stimuli are often highly simplified musical structures from western art music and thus lack both ecological validity and cross-stylistic applicability. Furthermore, the educational motivation behind many of these tests means that an individual's scores are generally correlated with the amount of formal musical training of the participants.

Regarding self-report questionnaires on musical expertise, experience and engagement (Cuddy, Balkwill, Peretz, & Holden, 2005; Ollen, 2006; Werner, Swope, & Heide, 2006; Chin & Rickard, 2010), we found that these instruments tend to rely on correct self-assessment by the participants and are thus prone to subjective bias. These tests are therefore unable to reveal musical abilities of which the participant is unaware and they also do not address abilities that the participant may be unable to verbalise. Also, while most of these self-report inventories cover a range of musical behaviours and types of engagement, they often do not comprehensively record the full range and level of musical skills and abilities, which is conversely a core goal of the Gold-MSI. Finally, some self-report inventories, such as Ollen's (2006) questionnaire, are only designed to measure abilities in special populations, such as trained musicians, and have only been validated on these populations.

2.1.2 Aims of the new musical sophistication index

Because the properties of the previously mentioned musical measurement instruments are not conducive to the assessment of musical sophistication as conceptualised below (see *The concept of musical sophistication*), we sought to develop a novel musical sophistication index for use with the general Western population. Our index combines verbal self-reports with quantitative measures of musical abilities in order to avoid as much as possible both a bias towards any specific musical style and also an excessive reliance on participants' self-assessment. Thus, we aim to assess and meaningfully combine both subjective self-reported musical behaviour and the objective performance on several music production and perception tests.

For the Goldsmiths Musical Sophistication Index our aim was to construct a measurement tool capable of assessing the multi-faceted profile of an individual's musical sophistication, which should be also be sensitive to facets of which participants are not themselves aware. To safeguard the applicability of the tool across the general population we incorporated aspects of musical behaviour which do not depend upon formal instrumental training; we took care to avoid the use of special musical terminology and we drew test items from a range of musical styles. The measurement instrument needs to be complemented by data norms based on a large sample from the general adult population of a western cultural sphere. These data norms can then be used to draw comparisons between and make relative judgements about individual performances.

2.1.3 The concept of musical sophistication

We felt the need to define musical sophistication as a concept for the evaluation of musical behaviour in order to distinguish it from terms such as musical ability, achievement, aptitude, audiation, engagement, experience, expertise, potential, skill, intelligence, talent, as well as musicality. These terms already come with defined meanings from the literature. Musical sophistication in our definition captures and combines important aspects from most of these existing concepts whilst still trying to set a novel focus. In our definition *Musical Sophistication* is a psychometric construct that comprises skills and achievements on a range of different and largely independent dimensions such as musical perception and music-making, amount of practice, emotional and functional usage of music, and creativity, where each dimension can be measured on a different subscale. As general principles of Musical sophistication we start with the following assumptions that:

- Facets of musical sophistication can develop through active engagement with music in its many different forms.
- Individuals vary by their level of sophistication on the different subscales.
- High levels of musical sophistication are generally characterised by
  - a higher frequency with which that musical skill or behaviour is exerted,
  - a greater ease, accuracy or effect of the musical behaviour when executed,
  - a greater and more varied repertoire of behaviour patterns associated with it.
- These differences in observable behaviour are related to a greatly differentiated system of cognitive categorisation and processing (where we explicitly adopt a very broad notion of cognition including implicit processing and embodied cognition that might be relevant to understand and describe certain aspects musical behaviour).

In addition to these axiomatic assumptions of musical sophistication we hypothesise the following:

- Musical sophistication might be style-independent. Musical behaviours as well as cognitive patterns of experts in different musical styles could be very similar.
- There might exist a general factor of musical sophistication that has an effect on all sub-facets.

Please note that these two hypotheses still need verification through empirical research.

In sum, according to our concept of musical sophistication, higher levels on this psychometric construct are assumed to correspond to *more frequent* and/or *more accurate* and/or *more differentiated* behaviours within a given facet of musical engagement.

2.1.4 Facets of musical sophistication and item writing

On the basis of the literature survey as well as the theoretical framework for musical sophistication given above, we initially posited five distinct hypothetical dimensions of musical sophistication:

1. Engagement, Motivation, Resource Allocation: the degree to which individuals prioritise music-related activities;

2. Music-making, performance, and improvisation and creativity: the degree of dedication to goal-oriented music-making;

3. Usage, flexibility, psychological functions: the degree of conscious use of music to alter emotional and mood states and degree to which individuals engage in and derive pleasure from incidental musical activities in everyday life;

4. Ability to verbalize musical experiences: the flexibility and richness of oral or written expression on music-related topics;

5. Musical ear, musical memory: skill level for production and perception of pitch and rhythm, and memorisation and recall of musical structures.

These five hypothetical dimensions served to orient the writing of the initial pool of question items. The following guiding principles were used for writing questionnaire items:

1. We tried to balance statements expressing a negation (e.g., "I would not consider myself a musician") with affirmative statements;

2. We aimed for questions that would apply, as much as possible, to any musical style and any age group, and paid attention to the specific vocabulary used;

3. We tried to be exhaustive and to cover all potential behaviours of interest for each dimension. The target population for responding to questions were adults with a range of levels of formal musical training (from no training up to professional level). We did not try to capture finer differences between professional musicians.

The first iteration of the survey comprised 153 question items. From this initial item pool, ambiguous items, quasi-synonymous items, items that did not fit with the overall concept of musical sophistication, and items that would potentially apply to only a very small subpopulation were eliminated. The remaining 111 items were then used in a pilot survey. For each of the five hypothetical dimensions we ensured that a sufficient number of items were stated positively.

We adopted the same seven-point scale for all items ranging from complete agreement to complete disagreement, as is standard for many psychological instruments. This scale includes a middle (i.e. neutral) category and represents a compromise between an interval scale providing data for subsequent parametric analyses and a manageable number of categories where each category retains a meaning that can be expressed verbally.

2.1.5 The pilot survey

The goals of the pilot survey were firstly to reduce the number of items in the pool of question items whilst ensuring that sufficient items remain to adequately index each latent dimension underlying distinct facets or dimensions (we will use these two terms interchangeably in the following). Secondly, we aimed to eliminate and adjust items such that for each dimension the range of levels of sophistication in the sample would be covered appropriately.

For the pilot survey, items were randomised and implemented in a simple online survey tool. The survey was made publicly available for one week in November 2010 and participants were mainly recruited through a link from the Science section of the BBC website (`http://www.bbc.co.uk/science/`). The pilot was taken by 488 participants (306 females and 182 males) with most participants (58%) in the age categories of 25-34 (32%) and 35-44 (26%) years with lower percentages of younger and older participants. 37% of the participants had completed an undergraduate degree as their highest level of education while lower proportions of the sample had achieved a postgraduate degree (29%), were still in education (17%) or had only completed A-levels (10%). 83% of the sample indicated the UK as their current country of residency while the other 17% were almost evenly spread among 26 other countries. Similarly, the most common country where participants had spent their formative years of childhood and youth was the UK (80%), again with a wide and almost even spread across a large number of other countries. 89% indicated "White" as their ethnic background.

We first scanned the distributions of answers on all 111 items for any items where individual answer categories were extremely over-represented. We only identified one item where this was the case (72% of all answers allocated to the lowest category and 97% fell into the lowest three categories). This item ("I don't know what people mean when they say I have a tune going round my head") was later removed from the item pool because of its low communality within the subsequent factor analyses.

The Kaiser-Meyer-Olkin measure of sampling adequacy suggested that the correlational structure of the data was very suitable for a factor analysis ($KMO = .931$) and we therefore analysed the data with several factor analyses using principal axis factoring but varying the number of dimensions to be extracted (from 5 dimensions, the initial number of dimensions for item writing, to 13 dimensions, which was indicated as an upper bound by a parallel analysis (Horn, 1965) comparing simulated random data to the actual data from the pilot). From the 13-factor solution we inspected item communalities and removed all items that would not fit even with the maximal factor solution, i.e. we removed all items with a communality of 0.4 or lower. This removed 11 items from the item pool. We then factor-analysed the data in a second iteration and reviewed each factor solution in terms of the number of items associated with each dimension, the eigenvalues of the factors, the communality of the items and the interpretability of solution after promax rotation. From this review we arrived at an 8-factor solution and removed items with communalities of 0.4 or less on the 8-factor solution. This resulted in the removal of a further 8 items. Factor 8 of this solution only comprised 2 very similar items asking for the building of playlists in itunes or similar programs/devices. This appeared conceptually too narrow to warrant a separate dimension of musical sophistication and we therefore excluded these two items from the item pool as well. This left us with 89 items which were subjected again to a factor analysis leading to the adoption of a 7-factor solution.

Because we deemed 89 items still too many for a self-report inventory, and in order to optimise the psychometric properties of the seven individual dimensions, we then combined analytic approaches from item response theory (IRT) and classic psychometric analysis to decide on item inclusion. For each dimension we applied the graded response model (Samejima, 1969) to identify those items that had a low overall information content or covered a very similar region in ability space, as well as those items which would increase Cronbach's alpha considerably when excluded from their respective dimension. In most instances, IRT analysis and Cronbach's alpha coincided on the same items which were therefore excluded. Eventually, we arrived at 70 items on seven dimensions, where each dimension showed very good psychometric indicators (values of Cronbach's alpha ranging between .693 and .921). We factor-analysed the data from the final pool of 70 items using principal axis factoring and promax rotation. The 7-factor solution explained 53.6% of the total variance and inter-correlations between the seven factors were found to be low to moderate in magnitude (range: .071 to .617).

Inspecting the item structure of each dimension gave rise to the interpretations and naming of the seven dimensions summarised in *Table 1*.

In addition to the 70 question items we added one question asking about the primary instrument played (if any) which, as a categorical variable, does not correspond to any of the seven dimensions of the Gold-MSI.

The list of all 71 items and their association with the seven factors is given in the Excel file `Gold-MSIv09_All_Items_Scoring_Template.xls`.

In addition to the 71 questions that make up v0.9 of the the Gold-MSI as a psychometric instrument, the template of the paper copy (`Gold-MSIv09_paper_layout.pdf`) also includes 12 non-music related questions asking for basic demographics etc. These questions can be removed or replaced with proper demographic scales if necessary.

2.1.6 Version 1.0

Version 0.9 of the self-report questionnaire was derived from the analysis of the pilot survey. The resulting questionnaire with with 71 questions was then implemented as part of the large online survey *How Musical Are You?* by BBC LabUK. Over the course of 2011 this implementation gathered 147,633 valid responses. We used these data in two ways: A) to derive data norms for version 0.9 based on this large sample and b) to derive an optimal scale and sub-scale structure from this much larger dataset to eventually arrive at version 1.0 of the self-report questionnaire. The details of this scale development are reported in (Müllensiefen, Gingras, Stewart, & Musil, n.d.) and it suffices to say here that exploratory as well as confirmatory factor analysis were used to identify a hierarchical factor structure comprising 5 sub-factors as well as one general factor of musical sophistication that was indicated by the high inter-factor correlations as well as an analysis using MacDonald's ω for detecting hierarchical factor structures. After

Table 1: Summary of dimensions measured by the Gold-MSI v0.9.

Number	Short Name	Interpretation	Cronbach's alpha	Number of items
1	Importance	Importance of music in everyday life	.914	15
2	Perception & Production	Self-reported musical perception and production abilities	.731	16
3	Musical training	Life history of formal musical training	.922	9
4	Emotion	Importance of music for psychological (esp. emotional) functions	.816	8
5	Body	Music and associated bodily movement	.826	7
6	Creativity	Musical creativity and ability to join into musical activities with others	.861	9
7	Openness	Attendance of cultural music events/openness to new music	.693	6

identifying and confirming the factor structure, we used item response analyses to reduce the number of items per factor but without compromising the reliability of the sub-scales.

The sub-scale structure of version 1.0 is very similar in many respects to version 0.9 and the 5+1 dimensions and their respective number of items are summarised in *Table 2*.

Note that the 18 items of the General Musical Sophistication factor are already part of other sub-scales and hence version 1.0 contains 38 items (plus the categorical question regarding the main instrument played and the same 12 non-music related questions as in version 0.9).

2.1.7 Scoring and data norms

Given that the Gold-MSI was deliberately constructed to comprise largely independent factors of musical sophistication as measured by distinct subscales, it follows that these subscales can be used independently of each other. Thus, if a researcher is only interested in e.g. the amount of musical training of a sample of participants, s/he can choose to administer only sub-scale 3.

Version 0.9 was derived from exploratory factor analysis without any notion of a hierarchical factor. This means that the sub-scales version 0.9 cannot be added together. In order to arrive at an "overall score of musical sophistication" version 1.0 of the questionnaire needs to be used.

All items are scored on the same 7-point scale and receive equal weights for scoring, which is known to make the scale more robust (Floyd & Widaman, 1995).

The different sub-scales comprise different numbers of items and therefore have different overall ranges. A comparison of individuals across different sub-scales is therefore most meaningful with reference to norm data. The tables *Table 11*, *Table 12*, and *Table 13* give the values for mean and standard deviation of each dimension as well as the boundaries for the percentiles for version 0.9 arising from the pilot sample (n=488) and the *How Musical Are You?* version (n=147,633) as well as for version 1.0 (n=147,633).

2.2 Usage

All materials for administration and scoring of the Gold-MSI are provided in the repository.

Table 2: Summary of dimensions measured by the Gold-MSI v1.0.

Number	Short Name	Cronbach's alpha	Number of items
1	Active Engagement	.872	9
2	Perceptual Abilities	.873	9
3	Musical Training	.903	7
4	Emotion	.791	6
5	Singing Abilities	.87	7
G	General Musical Sophistication	.926	18

For administering the full Gold-MSI as a paper copy the template in `Gold-MSIv10_Selfreport_Questionniare.pdf` should be printed and distributed to participants. Scores on the paper forms for individual participants can then be entered into individual Excel sheets in the Excel file `Gold-MSIv10_All_Items_Scoring_Template.xls`. Use column E of the individual participant sheet to enter the raw data. The data for negatively-phrased items is then automatically reversed by the spreadsheet and all data are copied to the `Master` sheet (sheet 1). Scores for the seven subscales are automatically calculated in columns CF to CL on this sheet. Please note that when the answer to a question consists of a number of numerical ranges (e.g. *number of years*), you should still enter the number from 1 to 7 which corresponds to the range (i.e. its order of presentation on the page), rather than the number or numbers which the answer represents.

The file `Gold-MSIv10_All_Items_Scoring_Template.xls` contains blank sheets for the data of 150 participants only. Scoring templates for more participants can be simply entered by duplicating the `Part_01` sheet and adding additional rows to the `Master` sheet (note: this requires duplicating and adjusting an Excel array formula to point to the correct participant sheet. Consult the Excel Help system for information on how to edit array formulae if in doubt.).

To use the Gold-MSI in an online survey, use the items in the order given on the second sheet of `Gold-MSIv10_All_Items_Scoring_Template.xls` (use columns `"No. in survey"` to order the items from 1 to 7). Remember to reverse the negatively-phrased items before calculating total scores for each subscale.

2.3 Giving feedback

In situations where there is not enough time to enter responses into the spreadsheet in situ (such as with online testing), it might be useful to implement automatic scoring. This should be a trivial task for a computerised implementation of the survey; a suitable scoring mechanism is also provided by some 'turnkey' online survey solutions. First, negatively loading questions should have the scale reversed (so that a response of 1 would score 7, 4 would score 4, 2 would score 6 and so on). The positive- versus negative-loading questions are identified on the provided scoring spreadsheets. Dimension totals are calculated by adding the scores—reversed as appropriate—from questions belonging to the given dimension; this can also be determined from the provided spreadsheet. Dividing a dimension total by 7 times the number of questions loading on that dimension gives the proportional score; multiplying this by 100 gives the percentage score.

3 The genre sorting task

3.1 Rationale

Gjerdingen and Perrott (2008) and Krumhansl (2010) showed that listeners can recognise the genre—or even the artist or title—of musical excerpts based on very short fragments of 400ms. We hypothesised that one aspect of musical sophistication may be correlated with the ability to extract relevant information from short musical excerpts and draw comparisons between these, as expertise has been associated with access to superior domain-specific knowledge (Hughson & Boakes, 2002) and domain-specific metacognitive skills (Veenman & Elshout, 1999). The ability to access relevant knowledge from briefly heard cues should therefore indicate a certain form of listening expertise. Because genre boundaries may be subjective and change over time (Gjerdingen & Perrott, 2008), we used the categories defined by Rentfrow and Gosling (2003), who identified four main musical meta-categories or styles: 1) reflective/complex, 2) intense/aggressive, 3) upbeat/conventional, 4) energetic/rhythmic. Rentfrow and Gosling (2006) further showed that preferences for specific musical categories were stable over time and associated with stable personality traits.

3.2 Versions 0.9, 0.91, and 1.0

3.2.1 Stimuli v0.9

Four musical genres were selected, one for each of the four categories identified by Rentfrow and Gosling (2003): *jazz* (reflective/complex), *rock* (intense/aggressive), *pop* (upbeat/conventional), and *hip-hop* (energetic/rhythmic). Additionally, following Krumhansl's (2010) finding that the approximate date of composition of a song could be identified fairly accurately from short excerpts, specific decades were selected for each genre: 1960-70s for jazz, 1970-80s for rock, 1990-2000 for pop and hiphop.

Exemplary songs for each of these genres were selected from the suggestions of prototypical songs given on the `allmusic.com` website. In order to avoid the recognition of specific songs, they were only selected if not in the top 100 all-time list (`http://www.billboard.com/specials/hot100/charts/top100-titles-00.shtml`) and never reached the top rank on the UK billboard (`http://www.chartstats.com/number1s.php?all=true`).

Fragments of 400 ms were chosen for each song using the following criteria:

- No human voice.
- At least two recognisable notes in the fragment.
- The fragment should represent, as much as possible, the maximal timbral diversity (maximum number of instruments) of the song. However, this rule was secondary to the two previous rules and was not always respected; overall, we aimed for a representative sound.
- When possible, the fragment should be part of a repeated section.

Excerpts were cut directly from .WAV files; the processing was done with the Audacity audio editor. A 20ms fade-in and fade-out was added and the peak amplitude of the excerpts normalised to 0dB. A list of the finalised excerpts is given in *Table 3*.

Table 3: Musical excerpts used in genre task v0.9.

No.	Genre	Artist	Original album	Song	Date	Begin min.sec.ms
1	Jazz	Nat Adderley	Work Song/Movin' Along	Work Song	1960	3.38.005
2	Jazz	Wayne Shorter	Speak No Evil	Speak No Evil	1964	7.06.700
3	Jazz	Eddie Palmieri	Azucar Pa' Ti (Sugar For You)	Azucar	1965	1.13.500
4	Jazz	Roy Ayers	Mystic Voyage	Evolution	1975	1.44.880
5	Rock	Little Feat	Feats' Don't Fail Me Now	Oh Atlanta	1974	3.07.045
6	Rock	The Tubes	The Completion Backward Principle	Talk To Ya Later	1981	2.17.220
7	Rock	Heart	Dreamboat Annie	Crazy On You	1976	3.48.900
8	Rock	Bad Company	Desolation Angels	Rock & Roll Fantasy	1979	1.23.100
9	Pop	Jessica Simpson	I Wanna Love You Forever	I Wanna Love You Forever	1999	3.40.900
10	Pop	Mandy Moore	So Real	So Real	1999	1.42.480
11	Pop	The Cardigans	Life	Carnival	1995	3.06.200
12	Pop	Ace of Base	The Sign	The Sign	1993	2.55.900
13	Hip-hop	Big Punisher	Capital Punishment	Still Not A Player	1998	2.41.900
14	Hip-hop	Cypress Hill	Black Sunday	I Ain't Goin' Out Like That	1993	3.45.000
15	Hip-hop	Eve	Scorpion	Who's That Girl	2001	4.00.550
16	Hip-hop	Public Enemy	Apocalypse 91: The Enemy Strikes Black	By The Time I Get To Arizona	1991	4.20.150

Table 4: Musical excerpt beginning times in genre task v0.91.

No.	Genre	Song	Begin A min.sec.ms	Begin B min.sec.ms
1	Jazz	Work Song	3.38.005	4.02.855
2	Jazz	Speak No Evil	0.05.505	7.06.700
3	Jazz	Azucar	1.13.500	5.32.500
4	Jazz	Evolution	1.44.880	1.49.975
5	Rock	Oh Atlanta	3.07.045	4.03.645
6	Rock	Talk To Ya Later	2.17.220	3.16.470
7	Rock	Crazy On You	0.55.000	3.48.900
8	Rock	Rock & Roll Fantasy	1.23.100	2.23.310
9	Pop	I Wanna Love You Forever	0.04.620	3.40.500
10	Pop	So Real	1.42.480	2.04.150
11	Pop	Carnival	3.06.200	3.10.385
12	Pop	The Sign	2.11.365	0.29.145
13	Hip-hop	Still Not A Player	2.41.900	3.08.300
14	Hip-hop	I Ain't Goin' Out Like That	3.45.000	3.56.670
15	Hip-hop	Who's That Girl	0.10.720	4.10.310
16	Hip-hop	By The Time I Get To Arizona	2.49.340	4.20.150

3.2.2 Procedure v0.9

Participants were instructed to sort the 16 excerpts by perceived similarity into 4 groups. The task was a constrained sorting problem: participants were told to include precisely 4 excerpts per group. Instructions (see *Appendix B*) stressed that they should sort by perceived similarity and the word *genre* was not mentioned. Excerpts were identified by icons on a computer screen, while groups corresponded to boxes. Participants could listen to an excerpt by hovering over its icon, and could move icons around by clicking and dragging. Participants could listen to each excerpt as many times as they wanted, in any order. There were no time constraints.

3.2.3 Stimuli v0.91

Pilot testing using the v0.9 stimuli suggested that task difficulty was quite high. Further excerpts were taken from the same songs, with the aim of including longer—and therefore easier—800ms items in the test (Gjerdingen & Perrott, 2008). For each song, an additional beginning-point for excerpts was identified (using the rules outlined previously), so that performance could be compared within-participants for different excerpt durations without a learning effect from repeated exposure. For four songs, the beginning times as used in the v0.9 stimuli included voice when stimulus duration was extended to 800ms, therefore alternative excerpts were identified from the same source songs. *Table 4* gives the additional excerpt beginning times identified for this test.

3.2.4 Procedure v0.91

The testing procedure was identical to that outlined for v0.9, above. However, after participants submitted their solution to one set of stimuli, they sorted a second set such that they completed one of each duration and one of each version (in a random order). Therefore, participants could sort one of:

- 400ms version A + 800ms version B,
- 800ms version A + 400ms version B,
- 400ms version B + 800ms version A,
- 800ms version B + 400ms version A.

3.2.5 Stimuli v1.0

Testing with the four v0.91 stimulus sets established that 800ms set B (completed by 131 participants) was the easiest (see *Table 4*). The floor effect of low scores observed for 400ms stimuli (see *Table 14*) was significantly reduced by using the 800b set; these 16 stimuli therefore form version 1.0 of the test. More detailed test statistics, as well as preliminary investigations of perceptual test attributes, are given in Musil, Elnusairi, & Müllensiefen, 2012.

3.2.6 Scoring versions 0.9 and 1.0

Two performance measures are calculated from individuals' grouping solutions on the test. Data norms for these are given in *Table14*. We have implemented a convenience function in `R`, which will return these performance scores and is documented in *Appendix E*. A link to the function, example code, and guidance for scoring spreadsheets is given in *3.2.8*.

Total pairs This is the count of correct pairs in the sorting solution per participant. Each possible pair of clips in a participant's solution is scored correctly if they are included in the same classification group, and their genre is the same. There are $\binom{n}{k}$, i.e. $\binom{16}{2}$= 24 pairs. Note that due to the combinations of the score measure, not all score values from 0-24 are possible. The resulting score value is not standardised, therefore comparisons to similar tests with more or fewer items are harder.

Adjusted Rand index The adjusted Rand index is used to measure the similarity of a given cluster solution to the optimal solution after accounting for chance (where every cluster contains items which belong in the same genre): $\frac{Index-ExpectedIndex}{MaxIndex-ExpectedIndex}$, based on observed, expected and maximum Rand indices $\frac{a-b}{x}$, where a is the number of correct pairs within clusters, b is the number of incorrect pairs across clusters, and x is the total number of pairs across the whole solution. A score of zero corresponds to roughly chance-level performance. We make use of the adjusted Rand index provided by the `R` library `e1071` in our scoring function, which for this version of the test returns a value between the theoretical lower and upper bounds of -0.25 and 1 respectively.

3.2.7 Data norms v0.9

Data norms for all of the documented performance scores on this task are given in *Table 14*.

3.2.8 Usage versions 0.9 and 1.0

All materials for administration and scoring of the Gold-MSI tests are provided in the repository. Depending on how you access the repository, the audio files may come in an individual archive, or in a subdirectory of a larger archive—in each case ensure that you have the correct test version. In each case, a number of `.wav` files will be present, each of which is a single 400ms long clip. File names begin with the length of the clip in milliseconds, the offset from the beginning of the recording in `minutes_seconds_milliseconds`, the genre of the clip, and the name of the song; for example: `400-0_05_505-Jazz-speak_no_evil.wav`. This information should be kept in the file names where possible, as it can prevent confusion when sanity checking the implementation of your test. Instructions for the task are provided in *Appendix B*, however you should additionally be careful not to discuss, label, or advertise the test using the word "genre" explicitly. This may discourage participants who may feel that they are unfamiliar with genres of popular music, even if they may otherwise perform well by sorting based on perceptual similarity (as is requested in the instructions).

The presentation of stimuli must confirm to the following specification (an example of which is given in the screenshots in *Appendix B*:

- Stimuli are represented by visual icons with a uniform size and shape, which can be visually moved to specific locations in some easy way.
- Any visual differentiation of stimuli must have as little potential to bias responses as possible. Colours may be appropriate (as long as they vary randomly and not systematically by brightness or hue), however letters, numbers, or symbols are certainly not. If there is any chance of repeat testing, such attributes must be randomly assigned to clips for each participant and each test.

- Interacting with the visual icon should immediately play the associated clip, and participants can audition clips as many times as they wish (in any order).

- Icons must be randomly assigned to starting positions in the initial (ideally 4x4) visual matrix or list per participant and per test.

- Bins into which the items are sorted must be separate from the initial pool and have a uniform size and shape. These may be differentiated in any abstract way (e.g. by letters, numbers, or colours, however not by explicit genre names, for instance).

- Bins must contain exactly 4 items in the final solution, i.e. all items must be evenly distributed across all bins.

- Whilst sorting, participants must be able to move icons freely between bins and the starting array.

One of the attractive properties of the test is that visually grouping by clicking and dragging icons around on a computer screen is quick, fun, and intuitive. The best solution would therefore be to program such a solution, or to use an online survey system which supports group sorting[3]. However, those with limited time and resources must also be able to implement this test. A possible shortcut which we used for some stages of piloting is to use commonly available media-player software. Anonymisation of any kind of graphical identification or differentiation of the song stimuli must be carried out in any case. Although it may be tempting to convert to `.mp3` or another compressed format with meta-data, the perceptual importance of timbral information in these short clips requires that they remain as uncompressed raw wave files. It may also be possible to use a drag-and-drop macro in Microsoft Powerpoint or similar software, with a suitable on-click behaviour to play the sounds.

When recording participants' scoring solutions, assign each bin a unique number from 1 to 4. For each item, record the number of the bin into which it was eventually sorted.

Although custom scoring mechanisms are not difficult to implement, the recommended method of scoring test data involves passing a correctly-formatted spreadsheet to our general purpose scoring function in `R`, along with an optional specification of which methods to run (for very large datasets, it can save time to run only the parts you require). For scoring genre-sorting data, the function needs to know how many elements to expect in each equally-sized bin (in the case of v0.9 and v0.91, this value is 4), which is given by the argument `binsize`. The scoring function (`gmsi_test_scoring.r`) is available in the test repository. The `R` code given in *Appendix E* assumes that `.csv` files are being used. The order of columns in the spreadsheet is not important, but the following columns *must be present, with named headers as below*:

- `uid` - any character string or number which uniquely identifies the participant and remains the same for all trials coded for that participant.

- `clip.id` - an integer from `1` to `16`, identifying the clip as in column `No.` in *Table 3*.

- `bin` - an integer from `1` to `4`, giving the bin into which the given clip was sorted.

Using the `R` scoring function (see *Appendix E* for more information, e.g. how to import and export data):

```
genre.scored <- performance.score(melody.data, method=c('prs', 'rnd'), binsize=4)
```

Depending on the methods chosen, the output can contain the following columns:

- `par` - uniquely identifies each participant (as in the input file)

- `prs` - total correct pairs

- `rnd` - adjusted rand index

[3]E.g. In one of our implementations we used a stock javascript group sorting question type on our online survey solution (`http://qualtrics.com`), and made individual flash movies to load within each draggable element. Each of these played their respective clip on mouse-over.

3.2.9 Giving feedback

Useful feedback can be given even when the scoring method described in *3.2.6* cannot be implemented automatically. In the BBC LabUK implementation of v0.9, the participant's scoring solution was re-displayed with the genre identity of clips revealed. Where this is not possible, participants can still be told which genres were being sorted.

4 The melody memory task

4.1 Rationale

The assessment of melodic memory has been a core component of musical ability and achievement tests for a long time (e.g. Seashore, 1919; Bentley, 1966). Melodic memory is believed to be a cognitive ability that is fundamental for many forms of musical engagement such as playing an instrument from memory, sight reading, learning new tunes, and spontaneous music making in a group context.

Melodic memory has been well researched since the 1970s and findings with regard to musical expertise (commonly measured as the amount of formal musical training) are ambiguous (see e.g. Halpern & Bartlett, 2010 for a review). Thus, it seems that only certain types of melodic memory—as tested by certain experimental paradigms—can be explained partially by formal musical training, therefore other facets of musical sophistication might be explaining (or partly explaining) the individual differences found in melodic memory experiments. The paradigm used here was inspired by the studies of Cuddy and Lyons (1981), and Dowling and Bartlett (1981), each of which used unfamiliar melodies and found no or only very weak effects of musical training on memory performance.

4.2 Versions 0.9, 0.91, and 1.0

4.2.1 Stimuli v0.9

Following Halpern, Bartlett, and Dowling (1995), novel melodies were created by shuffling the distributions of pitch intervals and rhythmic tone durations of existing and widely known folk or popular melodies. This was done to ensure overall stylistic familiarity with the melodies for western listeners, while making sure that no participant could have heard any of the test items before. The popular tunes that the newly created melodies are based on are named in *Table 5*, along with the manipulations used to create variants of the melodies (all melodies and their variants are given in score notation in *Appendix D*).

Table 5: Stimuli used in melodic memory task v0.9.

No.	Original tune name	Filename of original recomposition	Filename of recomposed variant	Same or different	Key original	Key variant	Key distance (in fifths)	Contour	In-key/out-of-key	Modified notes
1	Deck the halls	1_Deck_original_F.mid	1_Deck_nearby_C_violate_inkey.mid	Different	F	C	1	Violation	In-key	2
2	Do Re Mi	2_Doremi_original_C.mid	2_Doremi_distant_C#_preserve_outkey.mid	Different	C	C♯	5	Preserved	Out-of key	2
3	God bless America	3_GodBless_original_Bb.mid	3_GodBless_distant_A_identical.mid	Same	B♭	A	5	N/A	N/A	0
4	Hark the herald angels	4_Hark_original_Eb.mid	4_Hark_distant_D_violate_outkey.mid	Different	E♭	D	5	Violation	Out-of key	2
5	For he's a jolly good fellow	5_Jolly_original_Ab.mid	5_Jolly_nearby_Eb_violate_outkey.mid	Different	A♭	E♭	1	Violation	Out-of key	2
6	London bridge is falling down	6_London_original_C#.mid	6_London_nearby_F#_preserve_inkey.mid	Different	C♯	F♯	1	Preserved	In-key	2
7	Mary had a little lamb	7_Mary_original_F#.mid	7_Mary_distant_F_identical.mid	Same	F♯	F	5	N/A	N/A	0
8	My country 'tis of thee	8_MyCountry_original_B.mid	8_MyCountry_distant_Bb_preserve_inkey.mid	Different	B	B♭	5	Preserved	In-key	2
9	The first noel	9_Noel_original_E.mid	9_Noel_nearby_B_identical.mid	Same	E	B	1	N/A	N/A	0
10	Row row row your boat	10_RowRow_original_A.mid	10_RowRow_nearby_E_identical.mid	Same	A	E	1	N/A	N/A	0
11	We wish you a merry christmas	11_WeWish_original_D.mid	11_WeWish_nearby_G_identical.mid	Same	D	G	1	N/A	N/A	0
12	Yankee daddy	12_YankeeDaddy_original_G.mid	12_YankeeDaddy_distant_Ab_identical.mid	Same	G	A♭	5	N/A	N/A	0

The novel melodies were created by rearranging the order of pitches/intervals together with their rhythmic values. The guiding principle for the re-arrangement of the melodies was that they needed to make musical sense (i.e. have sensible phrase structure, a musically meaningful ending, motivic repetition where possible etc.). Melodies were between 10 and 16 notes long and lasted between 4 and 9 seconds. For each of six of these novel melodies a variant was created that differed in two notes from the original novel melody. An informal pilot indicated that variants differing in only one note seemed to be too difficult even for experienced listeners. Two types of differences between melodies were considered: Differences in contour (contour violations) and whether out of key notes were used in the variant (out-of-key violations). The six variant melodies covered all four possible combinations of these differences: Contour violation + in-key notes, contour violation + out-of-key notes, contour preservation + in-key notes, and contour preservation + out-of-key notes. In choosing altered notes in the case of in-key contour violations, care was obviously taken to avoid the suggestion of unexpected modes. All variants were transposed to a different key to ensure that memory for a melody's interval structure and not for absolute pitch differences was tested.

4.2.2 Stimuli v0.91

A larger set of stimuli was created by the addition of 16 more novel melodies to the existing stimulus set. Variants were created for 6 of these new unfamiliar melodies, again recomposed from the original folk tunes as in the existing set. In order to explore a wider range of difficulty levels three variants differed in only one tone and three variants differed in three tones each. The same two types of differences (contour and in- versus out-of-key notes) were used for creating differences between original and variant. *Table 6* gives the information for these melodies and variants and they can be found in score notation in *Appendix D*.

Table 6: Stimuli used in melodic memory task v0.91.

No.	Original tune name	Filename of original recomposition	Filename of recomposed variant	Same or different	Key original	Key variant	Key distance (in fifths)	Contour	In-key/out-of-key	Modified notes
13	Deck the halls	13_PDeck_original_C.mid	13_PDeck_nearby_G_identical.mid	Same	C	G	1	N/A	N/A	0
14	Do Re Mi	14_PDoremi_original_Bb.mid	14_PDoremi_nearby_F_identical.mid	Same	B♭	F	1	N/A	N/A	0
15	God bless America	15_PGodBless_original_Eb.mid	15_PGodBless_nearby_Bb_violate_inkey.mid	Different	E♭	B♭	1	Violation	In-key	1
16	Hark the herald angels	16_PHark_original_Ab.mid	16_PHark_distant_identical_A.mid	Same	Ab	A	5	N/A	N/A	0
17	For he's a jolly good fellow	17_Pjolly_original_C#.mid	17_Pjolly_distant_D_identical.mid	Same	C♯	D	5	N/A	N/A	0
18	London bridge is falling down	18_original_F#.mid	18_Plondon_nearby_B_identical.mid	Same	F♯	B	1	N/A	N/A	0
19	Mary had a little lamb	19_PMaryLamb_original_B.mid	19_PMaryLamb_nearby_violate_outkey_F#.mid	Different	B	F♯	1	Violation	Out-of key	3
20	My country 'tis for thee	20_PMyCountry_original_E.mid	20_PMyCountry_distant_Eb_identical.mid	Same	E	E♭	5	N/A	N/A	0
21	The first noel	21_PNoel_original_A.mid	21_PNoel_distant_violate_outkey_Ab.mid	Different	A	A♭	5	Violation	Out-of key	1
22	Row row row your boat	22_Prowboat_original_D.mid	22_Prowboat_distant_preserve_outkey_C#.mid	Different	D	C♯	5	Preserved	Out-of key	3
23	We wish you a merry christmas	23_Pwewish_original_G.mid	23_Pwewish_nearby_preserve_inkey_C.mid	Different	G	C	1	Preserved	In-key	1
24	Yankee daddy	24_PYankeeDaddy_original_F.mid	24_PYankeeDaddy_distant_preserve_inkey_E.mid	Different	F	E	5	Preserved	In-key	3
25	Happy birthday	25_PHappyBirthday_original_C.mid	25_PHappyBirthday_nearby_G_identical.mid	Same	C	G	1	N/A	N/A	0
26	Jingle bells	26_PJingle_original_Bb.mid	26_PJingle_nearby_F_identical.mid	Same	B♭	F	1	N/A	N/A	0
27	Oh come all ye faithful	27_POCome_original_C#.mid	27_POCome_distant_violate_inkey_D.mid	Different	C♯	D	5	Violation	In-key	3
28	Old Macdonald had a farm	28_Poldmac_original_Ab.mid	28_Poldmac_distant_preserve_outkey_A.mid	Different	A♭	A	5	Preserved	Out-of key	1

4.2.3 Stimuli v1.0

A shorter set of stimuli was chosen following piloting with the full stimulus set (as in *Table 6*) and 150 participants. Data were fitted to a Rasch measurement model, which estimates both participant ability and also item difficulty. Test items which did not fit this model were excluded. Items whose estimated difficulty was redundant (i.e. too close to other items) were also iteratively excluded. The measurement model was refitted each time, until the raw score distribution showed ideal characteristics (peaking around 75% of the score range, with less than 20% below chance level). 13 items form the optimised short test, and their stimulus excerpts are given in *Table 7*.

Table 7: Stimuli used in melodic memory task v1.0.

No.	Original tune name	Filename of original recomposition	Filename of recomposed variant	Same or different	Key original	Key variant	Key distance (in fifths)	Contour	In-key/out-of-key	Modified notes
1	Deck the halls	1_Deck_original_F.mid	1_Deck_nearby_C_violate_inkey.mid	Different	F	C	1	Violation	In-key	2
6	London bridge is falling down	6_London_original_C#.mid	6_London_nearby_F#_preserve_inkey.mid	Different	C♯	F♯	1	Preserved	In-key	2
7	Mary had a little lamb	7_Mary_original_F#.mid	7_Mary_distant_F_identical.mid	Same	F♯	F	5	N/A	N/A	0
8	My country 'tis of thee	8_MyCountry_original_B.mid	8_MyCountry_distant_Bb_preserve_inkey.mid	Different	B	B♭	5	Preserved	In-key	2
11	We wish you a merry christmas	11_WeWish_original_D.mid	11_WeWish_nearby_G_identical.mid	Same	D	G	1	N/A	N/A	0
12	Yankee daddy	12_YankeeDaddy_original_G.mid	12_YankeeDaddy_distant_Ab_identical.mid	Same	G	A♭	5	N/A	N/A	0
13	Deck the halls	13_PDeck_original_C.mid	13_PDeck_nearby_G_identical.mid	Same	C	G	1	N/A	N/A	0
14	Do Re Mi	14_PDoremi_original_Bb.mid	14_PDoremi_nearby_F_identical.mid	Same	B♭	F	1	N/A	N/A	0
15	God bless America	15_PGodBless_original_Eb.mid	15_PGodBless_nearby_Bb_violate_inkey.mid	Different	E♭	B♭	1	Violation	In-key	1
20	My country 'tis for thee	20_PMyCountry_original_E.mid	20_PMyCountry_distant_Eb_identical.mid	Same	E	E♭	5	N/A	N/A	0
25	Happy birthday	25_PHappyBirthday_original_C.mid	25_PHappyBirthday_nearby_G_identical.mid	Same	C	G	1	N/A	N/A	0
26	Jingle bells	26_PJingle_original_Bb.mid	26_PJingle_nearby_F_identical.mid	Same	B♭	F	1	N/A	N/A	0
27	Oh come all ye faithful	27_POCome_original_C#.mid	27_POCome_distant_violate_inkey_D.mid	Different	C♯	D	5	Violation	In-key	3

4.2.4 Procedure versions 0.9, 0.91, and 1.0

A training example preceded testing, using a short and well-known excerpt from the main theme of Mozart's Eine kleine Nachtmusik (KV 525) with an identical (but transposed) version as well as a transposed but altered variant. Written instructions were given to ensure that participants understood the terms 'same' and 'different' with respect to melody, while ignoring the transposition. The identical and altered variants were clearly labelled as such in the training example. Instruction text is given in *Appendix B.4*. For each trial participants first listened to the original, novel melody, followed by a two second silent interval after which the transposed variant was played[4]. Immediately after the variant finished playing participants were presented with a decision screen asking firstly for a same/different judgement and subsequently for a confidence judgement on a 3-point scale (*I'm guessing / I think so / I'm totally sure*). There were no time constraints for the responses to the decision screen. Performed without breaks between trials, the entire procedure took about four minutes for v0.9 and eight minutes for v0.91.

4.2.5 Scoring versions 0.9, 0.91, and 1.0

A number of performance measures are calculated from individuals' performance on the test. Data norms for these are given in *Table15*. We have implemented a convenience function in R, which will return any combination of these performance scores and is documented in *Appendix E*. A link to the function, example code, and guidance for scoring spreadsheets is given in *4.2.7*.

Accuracy Accuracy is quick to calculate and easily understood as a proportion or percentage of the maximum score: $Accuracy = Items_{correct}/Items_{total}$. It does not distinguish between participants who are genuinely guessing and those who scam the test by giving the same judgement for each trial.

***d*' (d prime)** *d*' ("d prime") is a measure of sensitivity accounting for both correct identifications of melodic difference, and also false alarms: $d' = Z(TPR) - Z(FPR)$, where Z is a standard z-score function of the standard normal distribution, TPR is the true positive rate, and FPR is false positive rate. This discriminates between biased responses and random guessing. z-scores of 0 and 1 are equal to negative and positive infinity respectively and need to be adjusted. We use a rule of thumb value for adjustment of 1 divided by twice the number of trials, which is always smaller than the next largest rate difference. Our scoring function determines the adjustment value from the number of rows per participant, but for exotic applications will also take an optional argument specifying this value.

Area Under Curve (AUC) Area Under Curve (AUC) refers to the area under the Receiver Operating Characteristic (ROC) curve, a function of true positive rate and false positive rate along a varying discrimination threshold. It can be thought of as reflecting the degree to which increasing numbers of correct identifications do not attract increasing numbers of false alarm judgements. To account for confidence as well as classification, we coded a new predictor variable, where 'different' judgements with the lowest confidence category received a value of 1, 'same' judgements with the highest confidence category received a value of 6, and other combinations of judgement and confidence were interpolated between these two. The user-friendly performance scoring function provided by the R package ROCR returned the AUC value per-participant for vectors of the recoded predictor value and correct value.

4.2.6 Data norms v0.9

Scores below chance level were adjusted upwards to chance level, affecting accuracies below 0.5, *d*' below 0, and AUC below 0.5; below-chance level performance is not informative. Data norms for all of the documented performance scores on this task are given in *Table 15*.

[4]For technical reasons, the silent interval in our online implementation of v0.91 was determined by the loading time between two webpages carrying the original and variant melodic stimuli respectively; we will revert to a strict interval in future versions.

4.2.7 Usage versions 0.9 and 1.0

All materials for administration and scoring of the Gold-MSI tests are provided in the repository. Depending on how you access the repository, the audio files may come in an individual archive, or in a subdirectory of a larger archive—in each case ensure that you have the correct test version. In each case, a number of `.wav` files will be present. Each file consists of two melodies separated by a 4000ms silent period. File names start with an identification number corresponding to column `No.` in *Table 5*. Two example files are also present; the first example file, `mozart_o_4000_mozart_v1.wav`, demonstrates transposition of an identical test melody, whilst the second example file, `mozart_o_4000_mozart_v2.wav`, demonstrates transposition of a test melody containing violations of both contour and tonality. These two files must be played in order and suitable instructions given (see *Appendix B*), so that participants understand which two deviations they are supposed to classify.

After the initial instructions and examples, all files in the test set need to be played in a randomised order for each participant. Suitable methods include any custom scripts, computer media players, and hardware media players (e.g. CD players) with the facility to 'shuffle' a playlist (or disc) without repeating any tracks. Participants should ideally listen to stimuli through sensible speakers or headphones, and should be in a quiet, distraction-free environment. It is especially important to check understanding of the concept of transposition before testing commences. Each trial consists of playing the given stimulus (a pair of melodies separated with a fixed silence) and recording the participant's response. Please refer to the instruction script given in *Appendix B*, as a model for verbal or displayed instructions given to participants. Participants judge whether the second melody in each pair was melodically identical—ignoring transposition—to the preceding target melody, and it is important to record this value as `1` for 'same' and `0` for 'different' judgements if the `R` scoring function is to be used later. Participants also give the subjective confidence of their judgement, which is recorded as `1` for "I'm guessing", `2` for "I think so", and `3` for "I'm totally sure".

Although custom scoring mechanisms are not difficult to implement, the recommended method of scoring test data involves passing a correctly-formatted spreadsheet to our general purpose scoring function in `R`, along with an optional specification of which methods to run (for very large datasets, it can save time to run only the parts you require). The scoring function (`gmsi_test_scoring.r`) is available in the test repository. We recommend using a `.csv` file as input, which can be created with all popular spreadsheet software and provides sensible defaults for delimiting fields and quoted text. The `R` code given in *Appendix E* assumes that `.csv` files are being used. The order of columns in the spreadsheet is not important, but the following columns *must be present, with named headers as below*:

- `uid` - any character string or number which uniquely identifies the participant and remains the same for all trials coded for that participant.
- `judgement` - participant's judgement (`0` or `1`), see above.
- `confidence` - participant's reported confidence (`1`, `2`, or `3`), see above.
- `true.answer` - true answer value, `0` if 'different' and `1` if 'same'.

Using the `R` scoring function (see *Appendix E* for more information, e.g. how to import and export data):

```
melody.scored <- performance.score(melody.data, method=c('acc', 'dpr', 'auc'))
```

Depending on the methods chosen, the output can contain the following columns:

- `par` - uniquely identifies each participant (as in the input file)
- `acc` - proportional accuracy score
- `tpr` - true positive rate
- `fpr` - false positive rate
- `dpr` - *d'* score
- `auc` - area under curve

4.2.8 Giving feedback

Any of the previously described scores can be given as feedback to curious participants. Accuracy is self explanatory (and easy to calculate by hand), d' can be thought of as a measure of sensitivity, and AUC can be thought of as consistency of discrimination ability after also taking confidence judgements into account.

5 The beat alignment perception task

5.1 Rationale

Following the Beat Alignment Test of Iversen and Patel (2008), the Beat Perception test aimed to investigate beat-based processing ability in the general population. The test might also prove helpful in identifying individuals who have particular difficulties in this regard. The Beat Perception test, as the name implies, considers the perception of the beat, independently of the ability to produce a beat. Excerpts of real music were chosen from a copyright-free library in order to give the task high ecological validity. This was maintained for the response criterion by deriving ideal beat points from several averaged performances by an expert drummer tapping in time to the excerpts.

5.2 Versions 0.9, 0.91, and 1.0

Twelve instrumental excerpts were chosen from three distinct genres which differ both stylistically and in terms of instrumentation (*rock, jazz, pop orchestral*). To avoid licensing issues, these were obtained from a copyright-free music library (`http://www.audionetwork.com/`). For all three genres, we included an excerpt in triple time and stimulus durations and tempi were, on average, equal for all three genres.

5.2.1 Stimuli v0.9

Each of the twelve excerpts was rendered in two versions. In one version a beep-track was superimposed such that the beeps coincided exactly with the actual beats present in the excerpt. In the second version, the beep-track was off the beat, in one of two ways: either the beeps were consistently ahead or behind the actual beats of the excerpt (out of phase) or the beeps were faster or slower than the tempo of the excerpt, in which case the beep track started on the beat at the beginning of the musical excerpt. The size of the alteration differed (either 2, 10 or 17.5%). Appropriate alterations were chosen on an excerpt by excerpt basis following piloting to cover a useful spectrum of difficulty across excerpts. *Table 8* gives details of stimuli implementations for test v0.9, as used in the BBC LabUK project.

Table 8: Stimuli used in beat alignment perception test v0.9

No.	Name	Genre	Metre	Tempo	Alteration	Alteration Direction	Alteration Size %	Duration (sec)
6	Prime rib	Rock	4	104	tempo	faster	2	14.1
9	Switchblade	Rock	4	159	phase	ahead	17.5	12.9
7	Psychedelic space	Rock	3	143	phase	ahead	10	15.2
3	Four handed hedgehog	jazz	4	142	tempo	slower	2	15.3
8	Sassy Stomp	jazz	4	115	phase	ahead	10	13.6
1	Crazy	jazz	3	165	phase	ahead	17.5	16
2	Freedom of the city	pop/orch	4	132	tempo	slower	2	11.7
4	For king and country	pop/orch	4	85	phase	ahead	17.5	10.9
5	Lord Arbinger Waltz	pop/orch	3	165	phase	ahead	17.5	13.0
ex3	One jump ahead (example)	pop/orch	4	120	phase	ahead	17.5	12.0
ex2	Roaring twenties (example)	jazz	4	108	tempo	slower	10	13.5
ex1	Never going back again (example)	Rock	4	120	on beat	n/a	n/a	12.0

5.2.2 Procedure v0.9

For each trial, participants heard a single audio clip from the stimulus set. Half of the trials consisted of clips with a beep-track that was on the beat; the other half consisted of clips where the beep track was off the beat, either with a tempo or phase alteration, making a total of 18 trials. The order of trials was randomised differently for each participant. Participants were asked to decide if the beep-track was on the beat or not and to give their subjective confidence in the judgement on a three-point scale. Three practice trials were given beforehand, demonstrating each type of clip (beep track was on the beat; beep track was off the beat (tempo alteration); beep track was off the beat (phase alteration).

5.2.3 Stimuli v0.91

Each of the twelve excerpts was rendered in three versions. In one version a 'beep-track' was superimposed where the beeps coincided with the actual beats present in the excerpt. In the second version, the beep-track was consistently ahead or behind the actual beats of the excerpt (phase alteration). In the third version, the beep-track was faster or slower than the tempo of the excerpt. The size of each individual alteration (either 2, 10 or 17.5%) was chosen following piloting to assess difficulty on an excerpt by excerpt basis. *Table 9* gives details of stimuli implementations for test v0.9, as used in the BBC LabUK project.

Table 9: Stimuli used in beat alignment perception test v0.91

No.	Name	Genre	Metre	Tempo	Alteration	Shift Direction	Shift Size %	Duration (sec)
6	Prime rib	rock	4	104	tempo	faster	2	14.1
6	Prime rib	rock	4	104	phase	ahead	10	14.1
6	Prime rib	rock	4	104	none	-	-	14.1
9	Switchblade	rock	4	159	phase	ahead	17.5	12.9
7	Psychedelic space	rock	3	143	tempo	faster	2	15.2
7	Psychedelic space	rock	3	143	none	-	-	15.2
3	Four handed hedgehog	jazz	4	142	tempo	slower	2	15.3
3	Four handed hedgehog	jazz	4	142	phase	ahead	17.5	15.3
3	Four handed hedgehog	jazz	4	142	none	-	-	15.3
8	Sassy Stomp	jazz	4	115	tempo	slower	2	13.6
8	Sassy Stomp	jazz	4	115	phase	ahead	10	13.6
1	Crazy	jazz	3	165	tempo	faster	2	16
2	Freedom of the city	pop/orch	4	132	tempo	slower	2	11.7
2	Freedom of the city	pop/orch	4	132	phase	ahead	17.5	11.7
2	Freedom of the city	pop/orch	4	132	none	-	-	11.7
4	For king and country	pop/orch	4	85	tempo	slower	2	10.9
5	Lord Arbinger Waltz	pop/orch	3	165	tempo	faster	2	13.0
ex3	One jump ahead (example)	pop/orch	4	120	phase	ahead	17.5	12.0
ex2	Roaring twenties (example)	jazz	4	108	tempo	slower	10	13.5
ex1	Never going back again (example)	rock	4	120	on beat	n/a	n/a	12.0

5.2.4 Procedure v0.91

For each trial, participants heard a pair of audio clips. Within a pair of sound clips, the identity of the excerpt was the same, but one of the clips was heard with the beep-track on the beat while the other was heard with the beep track off the beat. Within the whole experiment, each excerpt was presented with both types of alteration, making a total of 18 trials. The order of trials was randomised for each participant and the order of clips (on or off the beat) within each pair was randomly counterbalanced in the stimulus set (i.e. across all participants). In a two-alternative forced choice, participants were asked whether the first or second clip within each pair had the beep track that was on the beat. Practice trials were given beforehand, demonstrating each type of clip (beep track was on the beat; beep track was off the beat (tempo alteration); beep track was off the beat (phase alteration).

5.2.5 Stimuli v1.0

A shorter set of stimuli was chosen following piloting with the full stimulus set (as in *Table 9*) and 102 participants, using a binary decision response criterion as in v0.9. Data were fitted to a Rasch measurement model, which estimates both participant ability and also item difficulty. Test items which did not fit this model were excluded. Items whose estimated difficulty was redundant (i.e. too close to other items) were also iteratively excluded. The measurement model was refitted each time, until the raw score distribution showed ideal characteristics (peaking around 75% of the score range, with less than 20% below chance level). 17 items form the optimised short test, and their stimulus excerpts are given in *Table 10*.

Table 10: Stimuli used in beat alignment perception test v1.0

No.	Name	Genre	Metre	Tempo	Alteration	Shift Direction	Shift Size %	Duration (sec)
6	Prime rib	rock	4	104	none	-	-	14.1
6	Prime rib	rock	4	104	tempo	faster	2	14.1
6	Prime rib	rock	4	104	phase	ahead	10	14.1
9	Switchblade	rock	4	159	phase	ahead	17.5	12.9
7	Psychedelic space	rock	3	143	none	-	-	15.2
7	Psychedelic space	rock	3	143	tempo	faster	2	15.2
8	Sassy Stomp	jazz	4	115	tempo	slower	2	13.6
8	Sassy Stomp	jazz	4	115	phase	ahead	10	13.6
3	Four handed hedgehog	jazz	4	142	none	-	-	15.3
3	Four handed hedgehog	jazz	4	142	tempo	slower	2	15.3
3	Four handed hedgehog	jazz	4	142	phase	ahead	17.5	15.3
2	Freedom of the city	pop/orch	4	132	none	-	-	11.7
2	Freedom of the city	pop/orch	4	132	tempo	slower	2	11.7
2	Freedom of the city	pop/orch	4	132	phase	ahead	17.5	11.7
4	For king and country	pop/orch	4	85	tempo	slower	2	10.9
5	Lord Arbinger Waltz	pop/orch	3	165	tempo	faster	2	13.0
1	Crazy	jazz	3	165	tempo	faster	2	16
ex3	One jump ahead (example)	pop/orch	4	120	phase	ahead	17.5	12.0
ex2	Roaring twenties (example)	jazz	4	108	tempo	slower	10	13.5
ex1	Never going back again (example)	rock	4	120	on beat	n/a	n/a	12.0

5.2.6 Procedure v1.0

The same binary discrimination task procedure was used as in v0.9 (see *5.2.2*), as piloting with a two-alternative forced-choice task (as in v0.91) found the task to be too easy.

5.2.7 Scoring versions 0.9, 0.91, and 1.0

A number of performance measures are calculated from individuals' performance on the test. Data norms for these are given in *Table16*. We have implemented a convenience function in `R`, which will return any combination of these performance scores and is documented in *Appendix E*. A link to the function, example code, and guidance for scoring spreadsheets is given in *5.2.9*. See 4.2.5 for a discussion of the performance measures *accuracy*, *d'*, and *AUC*.

5.2.8 Data norms v0.9

Scores below chance level were adjusted upwards to chance level, affecting accuracies below 0.5, *d'* below 0, and AUC below 0.5; below-chance level performance is not informative. Data norms for all of the documented performance scores on this task are given in *Table 16*.

5.2.9 Usage versions 0.9 and 1.0

All materials for administration and scoring of the Gold-MSI tests are provided in the repository. Depending on how you access the repository, the audio files may come in an individual archive, or in a subdirectory of a larger archive—in each case ensure that you have the correct test version. In each case, a number of `.mp3` files will be present. Each file contains one musical stimulus with overlaid beep track. File names start with a number corresponding to column `No.` in *Table 8*, followed by a word denoting whether the beep track is *on* or off (*stretched* or *shifted*). Three example files are also present; these are used to demonstrate a beep track which is on the beat (`ex1-on.mp3`), a beep track which is stretched (`ex2-stretch.mp3`), and a beep track which is shifted (`ex3-shift.mp3`). Before starting the test, these three files must be played in order and suitable instructions given (see *Appendix B*), so that participants understand which three kinds of beep track timing they should listen for.

After the initial instructions and examples, all files in the test need to be played in a randomised order for each participant. Suitable methods include any custom scripts, computer media players, and hardware media players (e.g. CD players) with the facility to 'shuffle' a playlist (or disc) without repeating any tracks. Participants should ideally listen to stimuli through sensible speakers or headphones, and should be in a quiet, distraction-free environment. Each trial consists of playing the given stimulus and recording the participant's response. Please refer to the instruction script given in *Appendix B*, as a model for verbal or displayed instructions given to participants. Participants judge whether the heard stimulus had a beep track which was in time with the beat of the music or not, and it is important to record this value as `1` for 'on the beat' and `0` for 'off the beat' if the `R` scoring function is to be used later. Participants also give the subjective confidence of their judgement, which is recorded as `1` for "I'm guessing", `2` for "I think so", and `3` for "I'm totally sure".

Although custom scoring mechanisms are not difficult to implement, the recommended method of scoring test data involves passing a correctly-formatted spreadsheet to our general purpose scoring function in `R`, along with an optional specification of which methods to run (for very large datasets, it can save time to run only the parts you require). The scoring function (`gmsi_test_scoring.r`) is available in the test repository. We recommend using a `.csv` file as input, which can be created with all popular spreadsheet software and provides sensible defaults for delimiting fields and quoted text. The `R` code given in *Appendix E* assumes that `.csv` files are being used. The order of columns in the spreadsheet is not important, but the following columns *must be present, with named headers as below*:

- `uid` - any character string or number which uniquely identifies the participant and remains the same for all trials coded for that participant.
- `judgement` - participant's judgement (`0` or `1`), see above.
- `confidence` - participant's reported confidence (`1`, `2`, or `3`), see above.
- `true.answer` - true answer value, `0` if 'off the beat' and `1` if 'on the beat'.

Using the `R` scoring function (see *Appendix E* for more information, e.g. how to import and export data):

```
bat.scored <- performance.score(bat.data, method=c('acc', 'dpr', 'auc'))
```

Depending on the methods chosen, the output can contain the following columns:

- `par` - uniquely identifies each participant (as in the input file)
- `acc` - proportional accuracy score
- `tpr` - true positive rate
- `fpr` - false positive rate
- `dpr` - *d'* score
- `auc` - area under curve

5.2.10 Giving feedback

Any of the previously described scores can be given as feedback to curious participants. Accuracy is self explanatory (and easy to calculate by hand), *d'* can be thought of as a measure of sensitivity, and AUC can be thought of as consistency of discrimination ability after also taking confidence judgements into account.

6 The beat alignment production task

6.1 Note

Please note that this test has so far only been run in one version (v0.9 as used by BBC Lab UK). Implementation information and data norms are in preparation at the time of writing.

6.2 Rationale

The rationale for this test is similar to that for the Beat Perception Task (see *5.1*). However, in contrast to the Beat Perception task, this task focused on production, as opposed to perception of the beat. Analogously to the perception and production of pitch where some individuals only have problems producing (not perceiving) pitch, it is possible that some individuals struggle to synchronise action with a beat despite successfully perceiving it. Thus different patterns of dissociation between perception and production are potentially possible.

6.3 Version 0.9

6.3.1 Stimuli v0.9

The stimuli were identical to those used in the Beat Perception Test (see *5.2*), but no beep tracks were included.

6.3.2 Procedure v0.9

For each trial, participants heard a single audio clip from the stimulus set and were required to tap in time to the beat on the spacebar of the computer keyboard. A blue disc pulsed on the screen to give visual feedback of the beat they were producing. There were a total of nine trials. Participants were not required to synchronise to a particular tactus level, allowing for the fact that different participants might synchronise to different levels of the metrical hierarchy. Participants practised with a single example sound clip that was not used in the main task. See *Appendix B* for instruction text.

References

Bentley, A. (1966). *Bentley measures of musical abilities.* London: Harrap.

Boyle, J. D., & Radocy, R. E. (1987). *Measurement and evaluation of musical experiences.* New York: Schirmer books.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Chin, T., & Rickard, N. (2010). The music USE (MUSE) questionnaire: An instrument to measure engagement in music. In *Proceedings of the 11th international conference on music perception and cognition* (pp. 1–4). Washington, USA: ICMPC 11.

Cuddy, L. L., Balkwill, L., Peretz, I., & Holden, R. R. (2005). Musical difficulties are rare: A study of "tone deafness" among university students. *Annals of the New York Academy of Sciences*, *1060*(1), 311–317.

Cuddy, L. L., & Lyons, H. I. (1981). Musical pattern recognition: A comparison of listening to and studying tonal structures and tonal ambiguities. *Psychomusicology: Music, Mind & Brain*, *1*(2), 15–33.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology: Music, Mind & Brain*, *1*(1), 34–49.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299.

Gjerdingen, R. O., & Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, *37*(2), 93–100.

Gordon, E. E. (1989). *Advance measures of music audiation.* Chicago: Riverside Publishing Company.

Halpern, A. R., & Bartlett, J. C. (2010). Memory for melodies. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception* (Vol. 36, pp. 233–258). Philadelphia, PA: Springer Science and Business Media LLC.

Halpern, A. R., Bartlett, J. C., & Dowling, W. J. (1995). Aging and experience in the recognition of musical transpositions. *Psychology and Aging*, *10*(3), 325–342.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hughson, A. L., & Boakes, R. A. (2002). The knowing nose: The role of knowledge in wine expertise. *Food Quality and Preference*, *13*(7-8), 463–472.

Hughson, A. L., & Boakes, R. A. (2009). Passive perceptual learning in relation to wine: Short-term recognition and verbal description. *The Quarterly Journal of Experimental Psychology*, *62*(1), 1–8.

Iversen, J. R., & Patel, A. D. (2008). The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In *Proceedings of the 10th international conference on music perception and cognition* (pp. 465–468). Sapporo, Japan: ICMPC 10.

Krumhansl, C. L. (2010). Plink: "Thin slices" of music. *Music Perception: An Interdisciplinary Journal*, *27*(5), 337–354.

Musil, J. J., Elnusairi, B., & Müllensiefen, D. (2012). Perceptual dimensions of short audio clips and corresponding timbre features. In *Proceedings of the 9th international symposium on computer music modelling and retrieval* (pp. 311–318). Queen Mary University of London, London, United Kingdom.

Müllensiefen, D., Gingras, B., Stewart, L., & Musil, J. J. (n.d.). *The Goldsmiths Musical Sophistication Index.* (in preparation)

Ollen, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings.* Doctoral thesis, Ohio State University, Ohio.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders. *Annals of the New York Academy of Sciences*, *999*(1), 58–75.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, *84*(6), 1236–1256.

Rentfrow, P. J., & Gosling, S. D. (2006). Message in a ballad. *Psychological Science*, *17*(3), 236–242.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Richmond, VA: Psychometric Society.

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

Seashore, C. E., Lewis, D., & Saetveit, J. G. (1960). *Seashore measures of musical talent.* New York: The Psychological Corporation.

Veenman, M., & Elshout, J. (1999, December). Changes in the relation between cognitive and metacognitive skills during the acquisition of expertise. *European Journal of Psychology of Education*, *14*(4), 509–523.

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The musical ear test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, *20*(3), 188–196.

Weiser, M., & Shertz, J. (1983). Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies*, *19*(4), 391–398.

Werner, P. D., Swope, A. J., & Heide, F. J. (2006). The music experience questionnaire: Development and correlates. *The Journal of Psychology*, *140*(4), 329–345.

Wing, H. D. (1962). A revision of the "wing musical aptitude test". *Journal of Research in Music Education*, *10*(1), 39–46.

Yau, H. (1999). *A construction of criterion-referenced for badminton test battery.* National College of Physical Education and Sports, Taoyuan, Taiwan, ROC.

A Data norms

A.1 Self-report questionnaire norms, pilot study with n=488

Table 11: Self reported dimension percentiles and descriptive statistics from a pilot sample of 488 participants.

		F1 Importance	F2 Abilities	F3 Training	F4 Emotions	F5 Body	F6 Creativity	F7 Openness
Scale Min.		15	16	9	8	7	9	6
Scale Max.		105	112	63	56	49	63	42
Mean		74.2	81.4	26.5	47.4	39.6	39.7	29.5
SD		14.4	8.9	10.7	5.9	6.5	10.4	6.7
Percentiles	1	33	57	9	31	22	14	17
	2	43	62	9	34	25	17	20
	3	46	64	9	35	26	18	21
	4	47	66	10	36	27	20	22
	5	48	67	10	37	28	20	22
	6	49	68	10	38	28	22	23
	7	51	70	11	39	29	23	23
	8	52	70	11	39	30	24	23
	9	53	71	12	39	30	25	23
	10	55	73	12	39	30	26	24
	11	56	73	13	40	31	27	24
	12	56	74	13	40	31	27	24
	13	57	74	14	40	32	28	24
	14	57	75	14	41	32	29	24
	15	58	75	15	41	33	30	25
	16	60	76	15	41	33	30	25
	17	60	77	16	41	33	31	25
	18	61	77	16	42	34	31	25
	19	62	78	16	42	34	31	25
	20	62	78	17	43	34	31	26
	21	62	78	17	43	35	32	26
	22	63	79	17	43	35	32	26
	23	64	79	18	43	35	32	26
	24	64	79	18	44	35	33	26
	25	65	79	18	44	35	33	27
	26	65	80	19	44	36	33	27
	27	66	80	20	44	36	34	27
	28	66	80	20	44	36	34	27
	29	67	80	20	45	37	34	27
	30	68	81	20	45	37	34	28
	31	68	81	21	45	37	35	28
	32	69	81	21	45	37	35	28
	33	70	81	22	45	37	35	28
	34	70	82	22	45	37	36	28
	35	70	82	23	46	38	36	28
	36	71	83	23	46	38	36	28
	37	71	83	23	46	38	36	28
	38	72	83	24	46	38	37	28
	39	72	84	24	46	38	37	29
	40	72	84	25	46	38	37	29
	41	72	85	25	46	39	38	29

Continues on next page

Continues from previous page

	F1 Importance	F2 Abilities	F3 Training	F4 Emotions	F5 Body	F6 Creativity	F7 Openness
42	73	85	26	47	39	38	29
43	73	86	26	47	39	38	29
44	73	86	26	47	39	38	29
45	74	87	27	47	39	39	29
46	74	87	27	47	40	39	29
47	75	88	28	47	40	39	29
48	75	88	28	48	40	40	29
49	75	89	28	48	40	40	29
50	76	89	29	48	41	40	30
51	76	89	30	48	41	40	30
52	77	89	30	49	41	41	30
53	77	89	30	49	41	41	30
54	77	90	31	49	41	41	30
55	78	91	31	49	41	41	30
56	78	91	31	49	42	41	30
57	78	91	31	49	42	42	30
58	79	91	32	50	42	42	30
59	79	92	32	50	42	42	30
60	79	92	33	50	42	43	31
61	80	92	34	50	42	43	31
62	80	93	34	50	43	43	31
63	80	93	34	50	43	43	31
64	81	93	35	50	43	44	31
65	81	93	35	51	43	44	31
66	81	94	35	51	43	44	32
67	82	94	36	51	43	44	32
68	83	94	36	51	43	45	32
69	83	94	36	51	44	45	32
70	83	95	36	51	44	45	32
71	84	95	37	52	44	46	32
72	84	95	37	52	45	46	33
73	84	96	37	52	45	47	33
74	85	96	38	52	45	47	33
75	85	97	38	52	45	47	33
76	85	97	38	52	45	47	33
77	86	97	39	53	45	48	33
78	86	97	40	53	45	48	33
79	86	98	40	53	46	49	33
80	87	98	41	53	46	49	34
81	88	98	41	53	46	50	34
82	88	98	41	53	46	50	34
83	88	99	42	54	46	50	34
84	89	99	42	54	46	51	34
85	89	99	42	54	47	51	35
86	89	100	42	54	47	51	35
87	90	100	43	54	47	52	35
88	90	100	44	54	47	52	35
89	91	100	44	54	47	53	35
90	91	101	44	55	47	53	35
91	92	102	45	55	48	54	36
92	92	103	45	55	48	54	36

Continues on next page

Continues from previous page

	F1 Importance	F2 Abilities	F3 Training	F4 Emotions	F5 Body	F6 Creativity	F7 Openness
93	94	104	46	55	48	55	36
94	94	104	47	56	48	56	36
95	95	105	48	56	49	56	37
96	96	105	48	56	49	57	37
97	97	106	49	56	49	58	38
98	99	106	49	56	49	60	38
99	101	106	50	56	49	62	39

A.2 Self-report questionnaire norms, v0.9

Table 12: Self reported dimension percentiles and descriptive statistics, from a sample of 147,633 participants.

		F1 Importance	F2 Abilities	F3 Training	F4 Emotions	F5 Body	F6 Creativity	F7 Openness
Scale Min.		15	16	9	8	7	9	6
Scale Max.		105	112	63	56	49	63	42
Mean		77.8	90.0	34.3	48.1	40.7	40.5	30.8
SD		15.2	13.9	14.7	6.1	6.4	11.1	5.8
Percentiles	1	35	51	9	31	22	13	15
	2	42	58	9	34	25	15	18
	3	45	61	9	35	26	18	19
	4	48	64	10	36	28	19	20
	5	50	66	11	37	29	21	21
	6	52	68	11	38	30	22	21
	7	54	69	12	39	30	23	22
	8	55	70	12	39	31	24	22
	9	56	71	13	40	32	25	23
	10	57	72	14	40	32	26	23
	11	58	73	14	40	33	26	23
	12	59	74	15	41	33	27	24
	13	60	75	15	41	33	28	24
	14	61	75	16	41	34	28	24
	15	62	76	16	42	34	29	25
	16	63	76	17	42	34	30	25
	17	63	77	17	42	35	30	25
	18	64	78	18	42	35	31	25
	19	65	78	18	43	35	31	26
	20	65	79	19	43	36	31	26
	21	66	79	20	43	36	32	26
	22	67	80	20	43	36	32	26
	23	67	80	21	44	36	33	27
	24	68	80	21	44	37	33	27
	25	68	81	22	44	37	33	27
	26	69	81	23	44	37	34	27
	27	69	82	23	45	37	34	27
	28	70	82	24	45	38	35	28
	29	70	83	24	45	38	35	28
	30	71	83	25	45	38	35	28
	31	71	83	25	45	38	36	28

Continues on next page

Continues from previous page

	F1 Importance	F2 Abilities	F3 Training	F4 Emotions	F5 Body	F6 Creativity	F7 Openness
32	72	84	26	46	38	36	28
33	72	84	26	46	39	36	29
34	73	85	27	46	39	36	29
35	73	85	27	46	39	37	29
36	74	85	28	46	39	37	29
37	74	86	28	47	39	37	29
38	75	86	29	47	40	38	29
39	75	87	30	47	40	38	30
40	75	87	30	47	40	38	30
41	76	87	31	47	40	39	30
42	76	88	31	47	40	39	30
43	77	88	31	48	41	39	30
44	77	89	32	48	41	39	30
45	78	89	32	48	41	40	30
46	78	89	33	48	41	40	31
47	78	90	33	48	41	40	31
48	79	90	34	49	41	40	31
49	79	91	34	49	42	41	31
50	80	91	35	49	42	41	31
51	80	91	35	49	42	41	31
52	80	92	36	49	42	42	32
53	81	92	36	49	42	42	32
54	81	92	37	50	42	42	32
55	81	93	37	50	43	42	32
56	82	93	38	50	43	43	32
57	82	94	38	50	43	43	32
58	83	94	39	50	43	43	32
59	83	94	39	50	43	43	33
60	83	95	39	51	43	44	33
61	84	95	40	51	43	44	33
62	84	96	40	51	44	44	33
63	85	96	41	51	44	45	33
64	85	96	41	51	44	45	33
65	85	97	42	51	44	45	33
66	86	97	42	52	44	46	34
67	86	97	43	52	44	46	34
68	87	98	43	52	45	46	34
69	87	98	44	52	45	46	34
70	87	99	44	52	45	47	34
71	88	99	44	52	45	47	34
72	88	99	45	53	45	47	35
73	89	100	45	53	45	48	35
74	89	100	46	53	46	48	35
75	89	101	46	53	46	48	35
76	90	101	47	53	46	49	35
77	90	102	47	53	46	49	35
78	91	102	48	54	46	50	36
79	91	102	48	54	46	50	36
80	91	103	49	54	47	50	36
81	92	103	49	54	47	51	36
82	92	104	50	54	47	51	36

Continues on next page

Continues from previous page

	F1 Importance	F2 Abilities	F3 Training	F4 Emotions	F5 Body	F6 Creativity	F7 Openness
83	93	104	50	54	47	51	37
84	93	105	51	55	47	52	37
85	94	105	51	55	47	52	37
86	94	106	52	55	47	53	37
87	95	106	52	55	48	53	37
88	95	107	53	55	48	54	38
89	96	107	54	55	48	54	38
90	96	108	54	56	48	55	38
91	97	108	55	56	48	55	38
92	97	109	55	56	49	56	39
93	98	109	56	56	49	57	39
94	99	110	57	56	49	57	39
95	99	110	57	56	49	58	40
96	100	111	58	56	49	59	40
97	101	112	59	56	49	60	41
98	102	112	60	56	49	61	41
99	104	112	61	56	49	62	42

A.3 Self-report questionnaire norms, v1.0

Table 13: Self reported dimension and general sophistication percentiles and descriptive statistics, from a sample of 147,633 participants.

		F1 Active Engagement	F2 Perceptual Abilities	F3 Musical Training	F4 Singing Abilities	F5 Emotions	F6 General Musical Sophistication
Scale Min.		9	9	7	7	6	18
Scale Max.		63	63	49	49	42	126
Mean		41.52	50.20	26.52	31.67	34.66	81.58
SD		10.36	7.86	11.44	8.72	5.04	20.62
Percentiles	1	15	29	7	9	21	32
	2	18	32	7	12	23	37
	3	20	35	7	13	24	41
	4	22	36	7	15	25	43
	5	23	37	8	16	26	46
	6	24	38	8	17	27	48
	7	25	39	9	18	27	50
	8	26	39	9	19	28	51
	9	27	40	10	19	28	53
	10	28	41	10	20	28	54
	11	28	41	11	21	29	55
	12	29	41	11	21	29	56
	13	29	42	11	22	29	57
	14	30	42	12	22	29	58
	15	30	43	12	23	30	59
	16	31	43	13	23	30	60
	17	31	43	13	23	30	61
	18	32	43	14	24	30	62
	19	32	44	14	24	30	63
	20	33	44	14	25	31	64
	21	33	44	15	25	31	64
	22	34	44	15	25	31	65
	23	34	45	16	26	31	66
	24	34	45	16	26	31	67
	25	35	45	17	26	31	67
	26	35	45	17	26	32	68
	27	35	45	18	27	32	69
	28	36	46	18	27	32	69
	29	36	46	19	27	32	70
	30	36	46	19	28	32	71
	31	37	46	20	28	32	71
	32	37	46	20	28	33	72
	33	37	47	21	28	33	73
	34	38	47	21	29	33	73
	35	38	47	21	29	33	74
	36	38	47	22	29	33	75
	37	39	48	22	29	33	75
	38	39	48	23	29	33	76
	39	39	48	23	30	34	76

Continues on next page

Continues from previous page

	F1 Active Engagement	F2 Perceptual Abilities	F3 Musical Training	F4 Singing Abilities	F5 Emotions	F6 General Musical Sophistication
40	39	48	23	30	34	77
41	40	48	24	30	34	77
42	40	49	24	30	34	78
43	40	49	25	31	34	79
44	41	49	25	31	34	79
45	41	49	25	31	34	80
46	41	49	26	31	34	80
47	41	50	26	31	35	81
48	42	50	26	32	35	81
49	42	50	27	32	35	82
50	42	50	27	32	35	82
51	42	51	28	32	35	83
52	43	51	28	33	35	84
53	43	51	28	33	35	84
54	43	51	29	33	36	85
55	43	51	29	33	36	85
56	44	52	29	33	36	86
57	44	52	30	34	36	86
58	44	52	30	34	36	87
59	45	52	30	34	36	87
60	45	53	31	34	36	88
61	45	53	31	34	37	89
62	45	53	31	35	37	89
63	46	53	32	35	37	90
64	46	53	32	35	37	90
65	46	54	32	35	37	91
66	46	54	33	36	37	91
67	47	54	33	36	37	92
68	47	54	33	36	38	93
69	47	55	34	36	38	93
70	48	55	34	37	38	94
71	48	55	34	37	38	94
72	48	55	35	37	38	95
73	48	56	35	37	38	96
74	49	56	35	38	38	96
75	49	56	36	38	39	97
76	49	56	36	38	39	98
77	50	57	36	38	39	98
78	50	57	37	39	39	99
79	50	57	37	39	39	100
80	51	58	38	39	39	100
81	51	58	38	40	40	101
82	51	58	38	40	40	102
83	52	58	39	40	40	103
84	52	59	39	41	40	103
85	53	59	40	41	40	104
86	53	59	40	41	40	105
87	53	60	40	42	40	106

Continues on next page

Continues from previous page

	F1 Active Engagement	F2 Perceptual Abilities	F3 Musical Training	F4 Singing Abilities	F5 Emotions	F6 General Musical Sophistication
88	54	60	41	42	41	107
89	54	60	41	43	41	107
90	55	61	42	43	41	108
91	55	61	42	43	41	109
92	56	61	43	44	42	110
93	56	62	43	44	42	111
94	57	62	44	45	42	113
95	57	62	44	45	42	114
96	58	63	45	46	42	115
97	59	63	46	47	42	117
98	60	63	46	48	42	118
99	61	63	47	49	42	121

A.4 Genre sorting norms, v0.9

Table 14: Percentile and descriptive data norms for the genre sorting task, from a sample of 138,469 participants.

		Total correct pairs	Adjusted Rand index
Scale Minimum		0	-0.25
Scale Maximum		24	1
Mean		7.356	0.133
Standard deviation		2.785	0.145
Percentiles	<=15	<=4	<=-0.042
	16–25	5	0.010
	26–43	6	0.063
	44–53	7	0.115
	54–70	8	0.167
	71–81	9	0.219
	82–88	10	0.271
	89–95	12	0.375
	96–98	13	0.427
	99	>=15	>=0.531

A.5 Melody memory norms, v0.9

Table 15: Percentile and descriptive data norms for the melodic memory task, from a sample of 138,469 participants.

		Accuracy	True positive rate	False positive rate	*d'*	Area Under Curve
Scale Minimum		0.5	0	0	0	0.5
Scale Maximum		1	1	1	3.463	1
Mean		0.748	0.718	0.243	1.561	0.791
Standard deviation		0.161	0.203	0.202	1.075	0.170
Percentiles	<=5	0.500	<=0.333	0.042	0.000	0.500
	6–10	0.500	0.333–0.500	0.042	0.000	0.500
	11–15	0.500–0.583	0.500	0.042	0.000–0.431	0.542–0.583
	16–20	0.583	0.500	0.042	0.431	0.597–0.639
	21–25	0.583–0.667	0.500–0.667	0.042	0.537–0.861	0.653–0.667
	26–30	0.667	0.667	0.042	0.861–0.967	0.681–0.708
	31–35	0.667	0.667	0.042–0.167	0.967	0.722
	36–40	0.667–0.750	0.667	0.167	0.967–1.398	0.750–0.764
	41–45	0.750	0.667	0.167	1.398	0.778–0.792
	46–50	0.750	0.667	0.167	1.398	0.806–0.819
	51–55	0.750	0.833	0.167	1.398–1.732	0.833
	56–60	0.750–0.833	0.833	0.167–0.333	1.732–1.935	0.833–0.861
	61–65	0.833	0.833	0.333	1.935	0.861–0.889
	66–70	0.833	0.833	0.333	2.162	0.889–0.903
	71–75	0.833	0.833	0.333	2.162	0.917
	76–80	0.833–0.917	0.833	0.333	2.162–2.699	0.917–0.944
	81–85	0.917	0.833–0.958	0.500	2.699	0.944–0.972
	86–90	0.917	0.958	0.500	2.699	0.972–1.000
	91–95	0.917–1.000	0.958	0.500–0.667	2.699–3.463	1.000
	>96	1.000	>0.958	>0.667	3.463	1.000

A.6 Beat alignment perception norms, v0.9

Table 16: Percentile and descriptive data norms for the BAT perception task, from a sample of 139,481 participants.

		Accuracy	True positive rate	False positive rate	*d'*	Area Under Curve
Scale Minimum		0.5	0	0	0	0.5
Scale Maximum		1	1	1	3.892	1
Mean		0.701	0.814	0.425	1.260	0.739
Standard deviation		0.123	0.152	0.173	0.800	0.134
Percentiles	<=5	0.500	<=0.556	<=0.111	0.000	0.500
	6–10	0.500–0.556	0.556	0.111–0.222	0.000–0.279	0.500–0.549
	11–15	0.556	0.556–0.667	0.222	0.279–0.291	0.556–0.586
	16–20	0.556–0.611	0.667	0.222–0.333	0.334–0.570	0.593–0.611
	21–25	0.611	0.667	0.333	0.570–0.625	0.617–0.642
	26–30	0.611	0.667	0.333	0.625–0.790	0.648–0.667
	31–35	0.611–0.667	0.778	0.333	0.861–0.904	0.667–0.685
	36–40	0.667	0.778	0.333	0.904–1.081	0.691–0.710
	41–45	0.667	0.778	0.333–0.444	1.081–1.195	0.716–0.722
	46–50	0.667–0.722	0.889	0.444	1.195	0.728–0.747
	51–55	0.722	0.889	0.444	1.301–1.360	0.747–0.765
	56–60	0.722	0.889	0.444	1.360–1.529	0.772–0.778
	61–65	0.722–0.778	0.889	0.444–0.556	1.529–1.592	0.784–0.796
	66–70	0.778	0.889–0.958	0.556	1.592–1.651	0.802–0.821
	71–75	0.778	0.958	0.556	1.651–1.871	0.827–0.833
	76–80	0.778–0.833	0.958	0.556	1.871–1.985	0.833–0.852
	81–85	0.833	0.958	0.556	1.985–2.162	0.858–0.877
	86–90	0.833	0.958	0.556–0.667	2.162	0.889–0.895
	91–95	0.833–0.889	0.958	0.667	2.162–2.496	0.907–0.938
	>96	>0.889	>0.958	>0.667	>2.496	>0.944

B Instruction text

Note Explanatory notes are given in *in italics*. Instruction text for task versions 1.0 largely repeats that used for version 0.91 and is not reproduced here. Note that the per-trial instruction for the beat perception test v1.0 follows the single binary discrimination format as given in v0.9 (rather than v0.91), and that quantities and timings in the instructions are adjusted to reflect the shorter v1.0 tests.

B.1 Briefing

B.1.1 Version 0.91

Thank you for agreeing to take part in this experiment

The Music, Mind and Brain group at Goldsmiths, University of London, is investigating the ways in which a broad range of people engage with music. Your participation in this experiment will help us to validate a questionnaire that measures various aspects of musicality.

You will be asked to complete two questionnaire tasks and three musical 'puzzles'. We hope that you will find these fun and stimulating! Please try to take this test in a quiet room; if possible, use headphones at a comfortable listening level. Computer speakers should also be fine. The whole task takes around 45 minutes and you may take breaks when you reach the instruction screens between tasks.

At the end of the experiment you will see your scores on some of the dimensions we are investigating. If you have any questions about the project (or your personal scores), please do not hesitate to get in touch with Jason Musil, at the email address ps001jm@gold.ac.uk

This experiment has been approved by the Ethics Committee at Goldsmiths College, University of London. Your data may be used in academic journals or selective press releases as part of our overall findings, however at no stage will any personally identifying data be made available. All data will be held in accordance with the Data Protection Act 1998 and destroyed within 5 years.

Please click CONTINUE to start.

B.2 Questionnaire instructions

B.2.1 Version 0.9

Your relationship with music
Please respond to the statements below:

Questions and possible responses are displayed across a number of pages.

B.2.2 Version 0.91

The Goldsmiths Musical Sophistication Index
This survey asks about your musical abilities and interests.

Click CONTINUE to start.

B.3 Genre sorting task instructions

B.3.1 Version 0.9

Instructions:
You will need your computer's sound turned on to do this test. This test assesses your ability to group

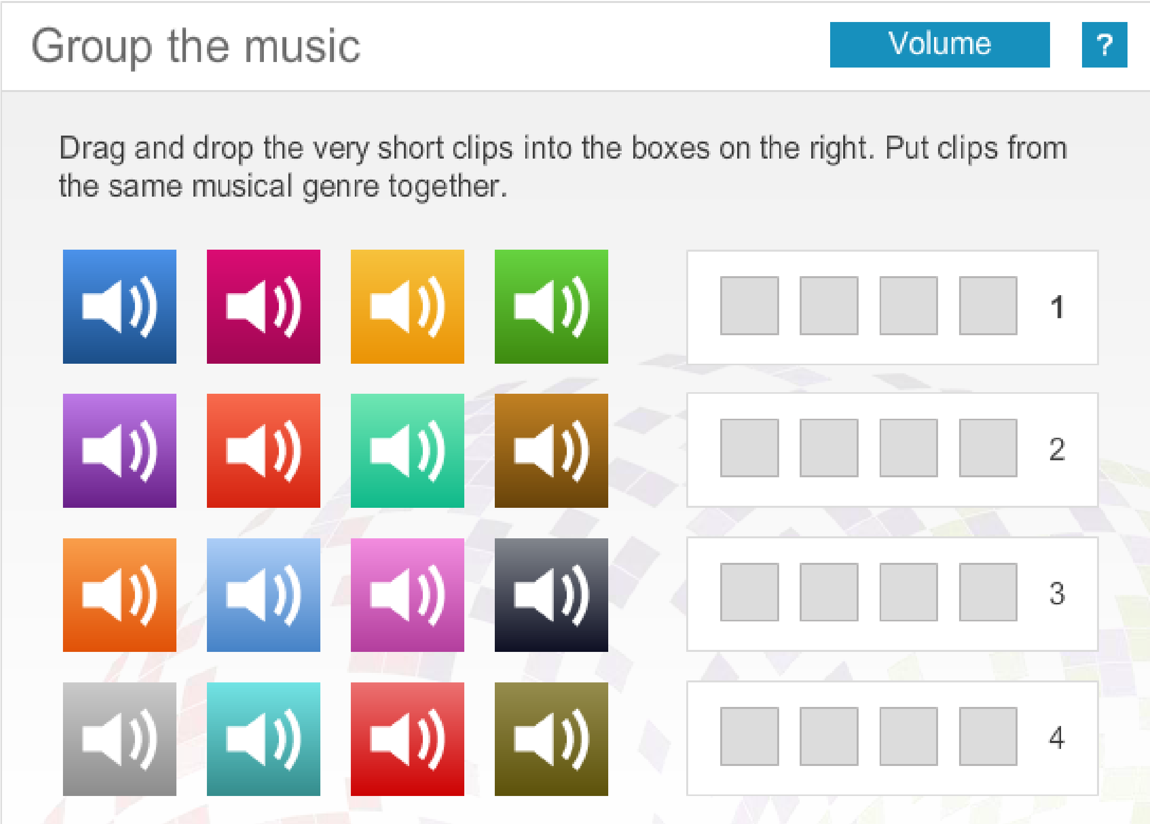


Figure 1: Task interface for genre sorting task.

together pieces of music based on how similar you think they sound[5]. There are sixteen VERY SHORT music clips, each represented at random by a different colored square. Click on a square to listen to a clip. Drag and drop the clips into the boxes on the right. Put clips of the same musical genre together. You can continue to drag clips in and out of the boxes until you are happy with your groups.

Click 'Continue' to start.

Task layout is displayed, see Figure B.3.1.

Drag and drop the clips into the boxes on the right. Put clips of the same musical genre together.

B.3.2 Version 0.91

Group the Music!
This test assesses your ability to group together pieces of music with a similar style. There are two sessions to complete. For each session there are 16 music clips, each represented by a blue square. You can listen to each clip by hovering over the square with your mouse. Drag and drop the clips into the boxes on the right. Put clips of the same musical style together. You can continue to drag clips in and out of the boxes until you are happy with your groups.

Please note that the clips may take a few moments to load at the start, please be patient.

Click CONTINUE to start.

[5]Please note that in the BBC LabUk implementation this sentence was mistakenly written: "This test assesses your ability to group together pieces of music from a similar genre (similar style)". The text supplied here is more correct.

Page changes to task layout.

Drag and drop the clips into the boxes on the right. Put clips of the same musical style together. You need to end up with four clips in each group.

B.4 Melody memory task instructions

B.4.1 Version 0.9

Instructions
This test assesses your memory for tunes. You will hear 12 pairs of tunes. The second tune in each pair will always be at a different "pitch" which means it is the same tune but played either higher or lower. Listen to each pair of tunes. Ignoring the difference in pitch, are the tunes exactly the same or slightly different? If exactly the same, answer SAME. If you think something has changed, answer DIFFERENT.

Click 'Continue' for a demonstration.

Example trials are given.

You will now hear 12 pairs of tunes. In each case decide if the second tune is the SAME or DIFFERENT to the first.

Remember to ignore the fact that the second tune in each pair will always be at a different pitch (which means the same but played higher or lower).

Click 'Continue' to start.

Task layout is displayed, see Figure B.4.1

Listen to this tune: *Stimulus is played and page changes.*

Now listen to this tune: *Stimulus is played and page changes.*

Was the second tune the same melody as the first, despite being played at a different pitch?

How sure are you about this answer?

B.4.2 Version 0.91

Test Your Melodic Memory!
This tests assesses your memory for tunes.

You will hear pairs of tunes. The second tune in each pair will always be at a different 'pitch', which means it is the same tune but played either higher or lower. Listen to each pair of tunes. Ignoring the difference in pitch, are the tunes exactly the same or slightly different?

If exactly the same, answer SAME. If you think something has changed, answer DIFFERENT.

You will now hear some examples. Each example will load automatically after the previous one has played. Please pay careful attention to the description text with each example.

Click CONTINUE for an example.

Page changes and stimulus is played.

This is the original tune.

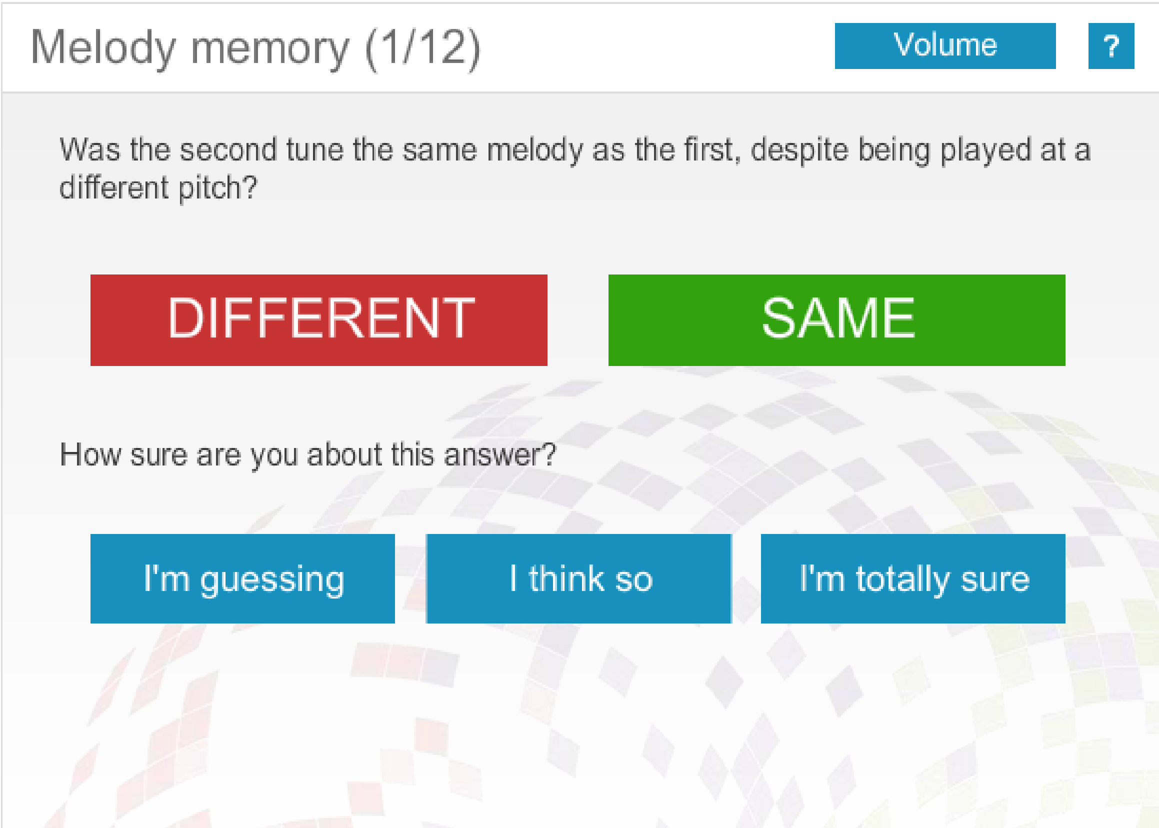


Figure 2: Task interface for melody memory task.

Page changes and stimulus is played.

This is the SAME melody as the original tune, but played lower in pitch.

Page changes and stimulus is played.

This is the original tune again.

Page changes and stimulus is played.

This tune contains a slight difference in the melody compared to the original tune, as well as being at a lower pitch. The correct answer would be DIFFERENT.

Page changes.

Instructions

You will now hear 28 pairs of tunes. In each case decide whether the second tune is the SAME or DIFFERENT to the first.

Remember to ignore the fact that the second tune in each pair will always be at a different pitch (which means it is overall higher or lower). Each tune will load automatically after the previous one has played.

Click CONTINUE to start.

Page changes and stimulus is played.

Listen to this tune.

Page changes and stimulus is played.

Now listen to this tune.

Page changes.

Was the second tune the same melody as the first, despite being played at a different pitch?

How sure are you about this answer?

B.5 Beat alignment perception task instructions

B.5.1 Version 0.9

Match the Beat - Instructions
You will need your computer's sound turned on to do this test. This test assesses your ability to recognise the beat in a piece of music. Listen to the following music clips. In each clip you will hear some music together with a "beep-track". Your task is to decide whether the "beep-track" is ON or OFF the beat.

Click 'Continue' to start.

Example trials are displayed, see Figure B.5.1.

In this example the beeps are ON the beat of the music. The correct answer would be "on the beat".

Stimulus plays and page changes.

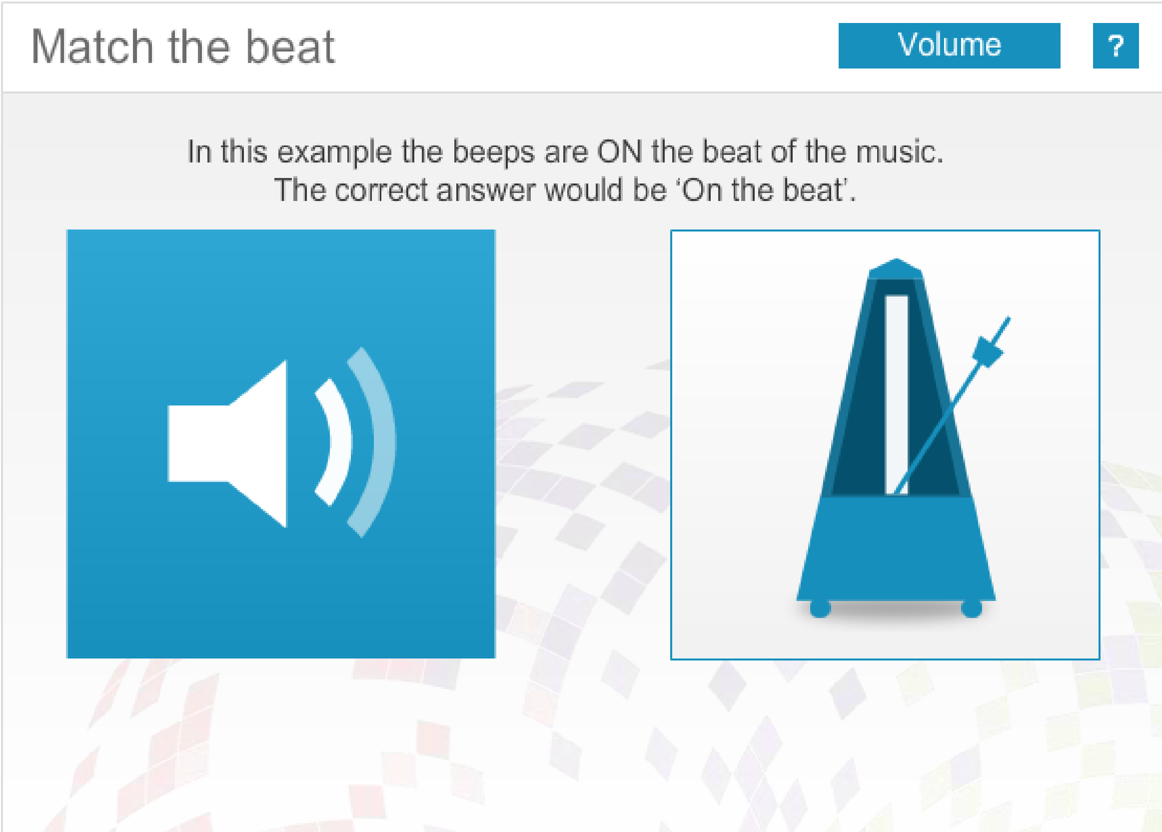


Figure 3: Task interface for beat alignment perception examples.

In this example the beeps are OFF the beat of the music. The correct answer would be "off the beat".

Stimulus plays and page changes.

In this example the beeps are OFF the beat of the music. The correct answer would be "off the beat".

Stimulus plays and page changes.

You will now hear 18 clips. In each case you will be asked to decide whether the beeps are ON or OFF the beat.

Click 'Continue' to start.

Task layout is displayed, see Figure B.5.1.

B.5.2 Version 0.91

Find the Beat!
This test assesses your ability to recognize the beat in a piece of music. Listen to the following music clips. In each clip you will hear some music together with a 'beep-track'. The beep-track can be ON or OFF the beat of the music.

You will now hear some examples. Each example will load automatically after the previous one has

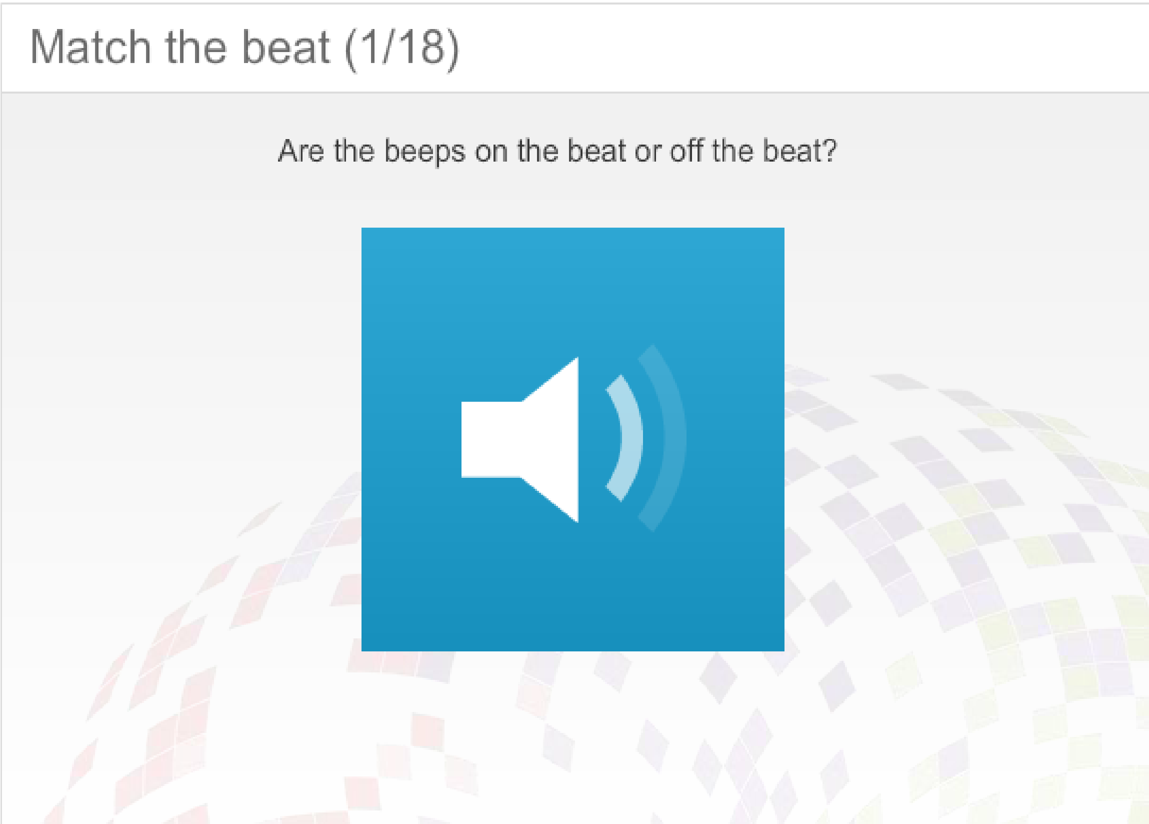


Figure 4: Task interface for beat alignment perception task.

played. Please pay careful attention to the description text with each example.

Click CONTINUE for an example.

Page changes and stimulus plays.

In this example the beep-track is ON the beat of the music.

Page changes and stimulus plays.

In this example the beep-track is OFF the beat of the music.

Page changes and stimulus plays.

In this example the beep-track is also OFF the beat of the music.

Page changes.

Instructions

You will now hear 18 pairs of music clips. In each case, decide whether the FIRST or the SECOND clip contains the beep-track that is ON the beat. Note: a musical clip may appear in more than one pair.

The second clip in each pair will load automatically after the previous one has played.

Click CONTINUE to start.

Page changes and stimulus plays.

Listen to this clip.

Page changes and stimulus plays.

Now listen to this clip.

Page changes.

Which of the two clips had a beep-track that was ON the beat?

How sure are you about this answer?

B.6 Beat alignment production task instructions

B.6.1 Version 0.9

Tap to the Beat - Instructions
This test assesses your ability to tap out the beat in a piece of music. Listen to the following music clips. As soon as you pick up the beat, use the spacebar on your keyboard to tap regularly in time with the beat until the clip ends. The blue circle will "pulse" when you tap so you can tell your spacebar taps are being registered. The system records your taps very accurately but depending on your computer there may be a small delay in showing the blue circle "pulse".

Click 'Continue' to have a practice.

Task layout is displayed, see Figure B.6.1.

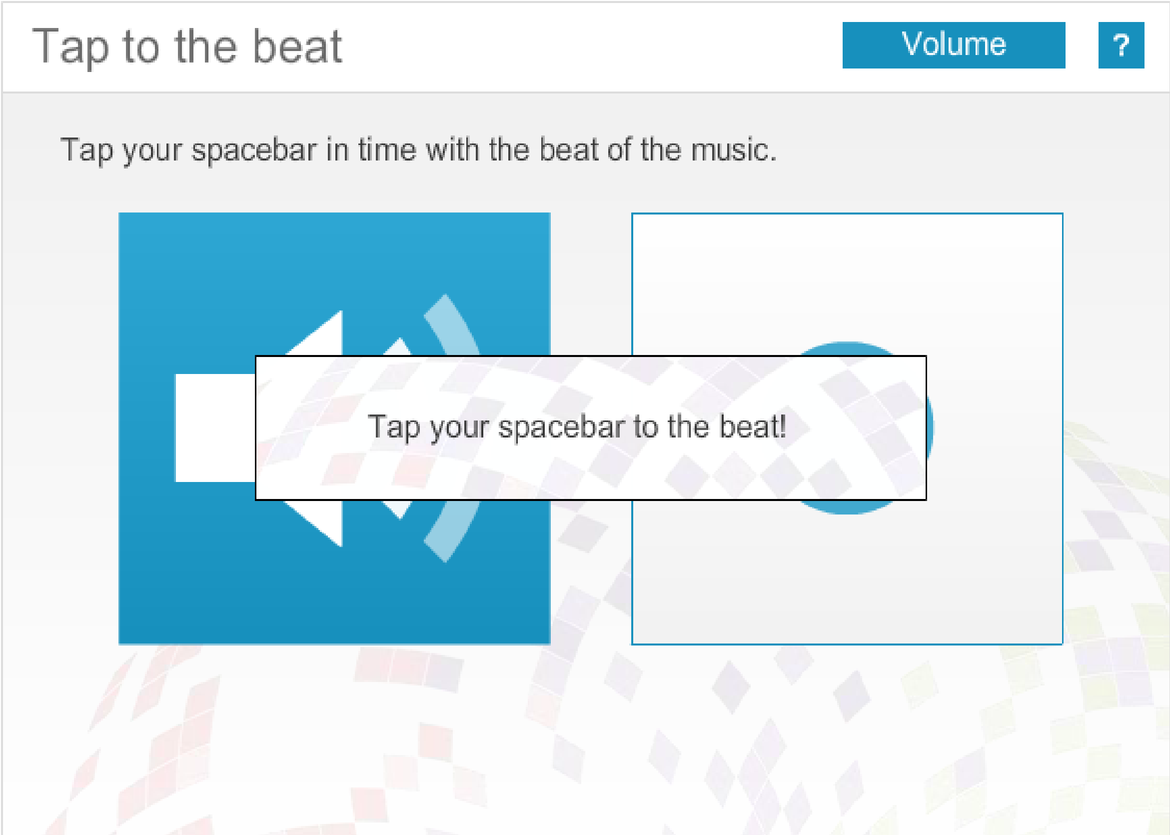


Figure 5: Task interface for beat alignment production task.

Tap your spacebar in time with the beat of the music.

B.6.2 Version 0.91

Not implemented.

C Links to original sound files

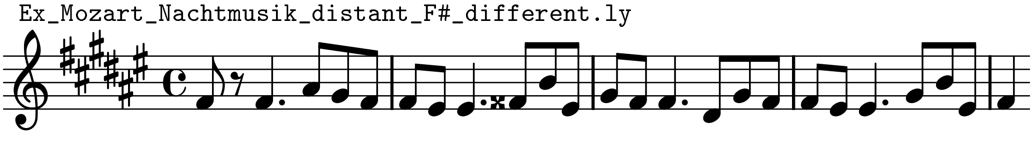
C.1 Beat alignment test stimuli v0.9 and v0.91

- Prime Rib
  http://www.audionetwork.com/production-music/prime-rib_15643.aspx
- Never Going Back Again
  http://www.audionetwork.com/production-music/never-going-back-again_17255.aspx
- Psychedelic Space
  http://www.audionetwork.com/production-music/psychedelic-space_9831.aspx
- Four Handed Hedgehog
  http://www.audionetwork.com/production-music/four-handed-hedgehog_33902.aspx
- Sassy Stomp
  http://www.audionetwork.com/production-music/sassy-stomp_26241.aspx
- Crazy 2
  http://www.audionetwork.com/production-music/crazy-2_5198.aspx
- One Jump Ahead
  http://www.audionetwork.com/production-music/one-jump-ahead_35559.aspx
- Freedom Of The City
  http://www.audionetwork.com/production-music/freedom-of-the-city_35540.aspx
- Roaring Twenties
  http://www.audionetwork.com/production-music/roaring-twenties_26259.aspx
- For King And Country
  http://www.audionetwork.com/production-music/for-king-and-country_29632.aspx
- Lord Abinger Waltz
  http://www.audionetwork.com/production-music/lord-abinger-waltz_19351.aspx
- Switchblade 2
  http://www.audionetwork.com/production-music/switchblade-2_14006.aspx

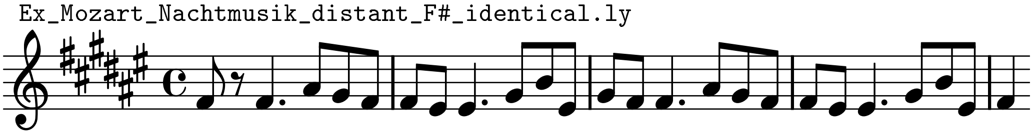
D Melodic stimuli

D.1 Example stimuli

Ex_Mozart_Nachtmusik_distant_F#_different.ly



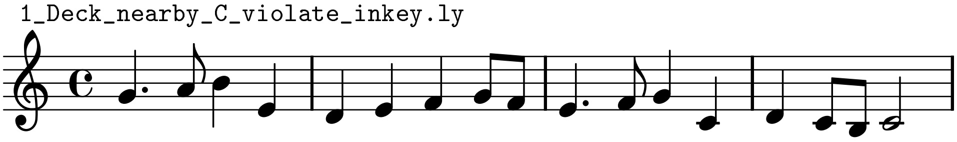
Ex_Mozart_Nachtmusik_distant_F#_identical.ly



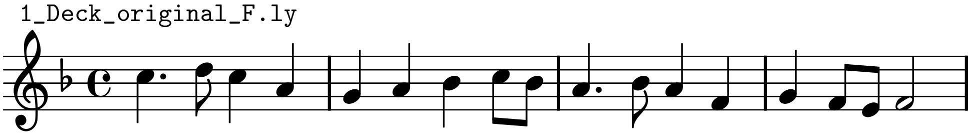
Ex_Mozart_Nachtmusik_original_G.ly

D.2 Stimuli v0.9

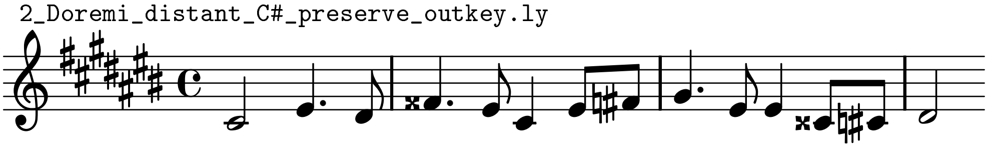
1_Deck_nearby_C_violate_inkey.ly



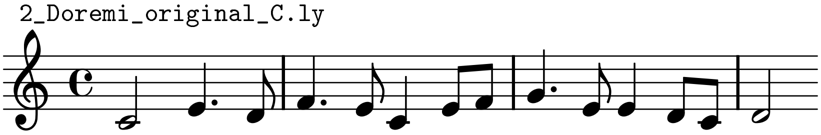
1_Deck_original_F.ly



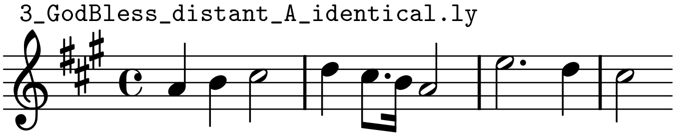
2_Doremi_distant_C#_preserve_outkey.ly



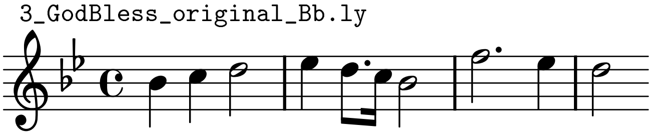
2_Doremi_original_C.ly



3_GodBless_distant_A_identical.ly



3_GodBless_original_Bb.ly

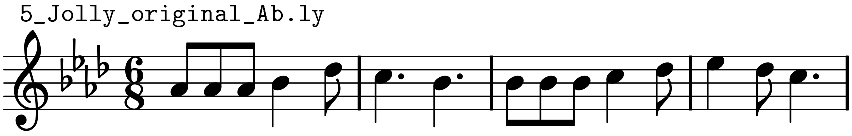


4_Hark_distant_D_violate_outkey.ly

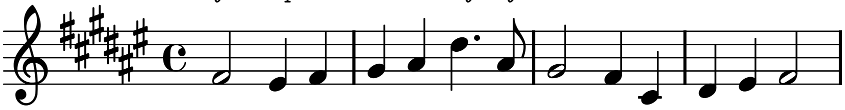
4_Hark_original_Eb.ly

5_Jolly_nearby_Eb_violate_outkey.ly

5_Jolly_original_Ab.ly

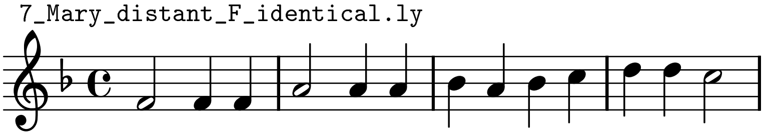


6_London_nearby_F#_preserve_inkey.ly

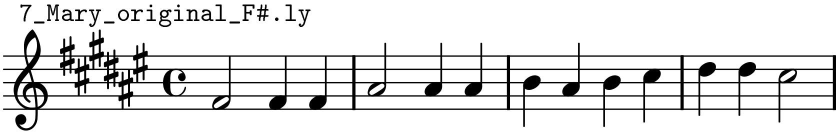


6_London_original_C#.ly

7_Mary_distant_F_identical.ly



7_Mary_original_F#.ly



8_Mycountry_distant_Bb_preserve_inkey.ly

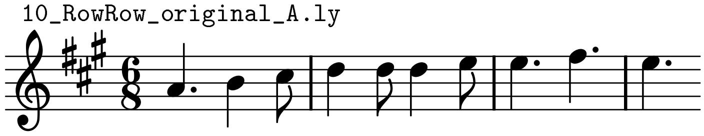
8_Mycountry_original_B.ly

9_Noel_nearby_B_identical.ly

9_Noel_original_E.ly

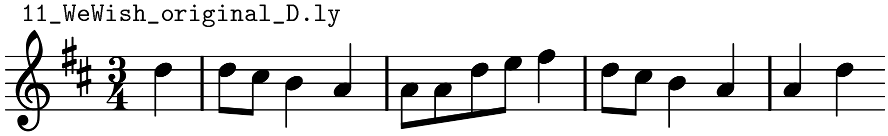
10_RowRow_nearby_E_identical.ly

10_RowRow_original_A.ly

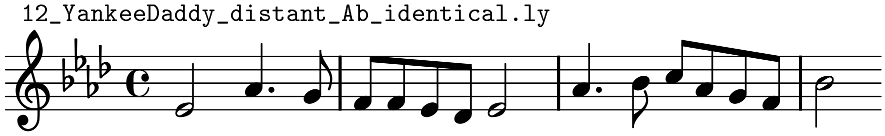


11_WeWish_nearby_G_identical.ly

11_WeWish_original_D.ly



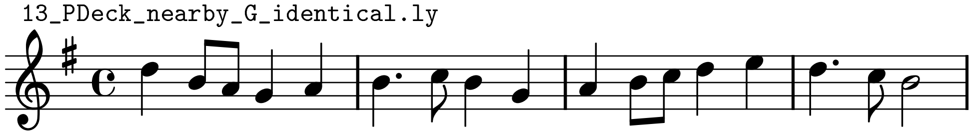
12_YankeeDaddy_distant_Ab_identical.ly



12_YankeeDaddy_original_G.ly

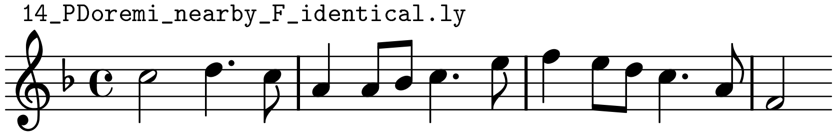
D.3 Stimuli v0.91

13_PDeck_nearby_G_identical.ly



13_PDeck_original_C.ly

14_PDoremi_nearby_F_identical.ly

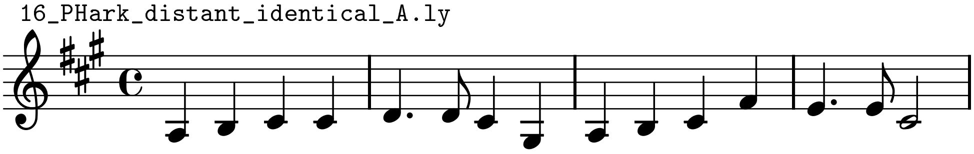


14_PDoremi_original_Bb.ly

15_PGodBless_nearby_Bb_violate_inkey.ly

15_PGodBless_original_Eb.ly

16_PHark_distant_identical_A.ly



16_PHark_original_Ab.ly

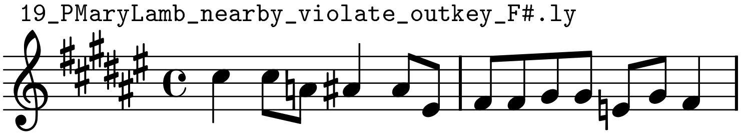
17_Pjolly_distant_D_identical.ly

17_Pjolly_original_C#.ly

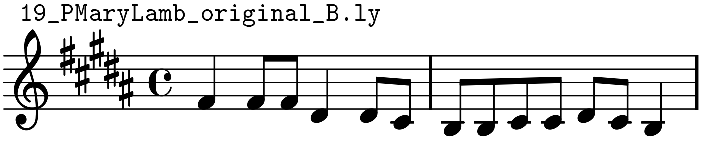
18_Plondon_nearby_B_identical.ly

18_Plondon_original_F#.ly

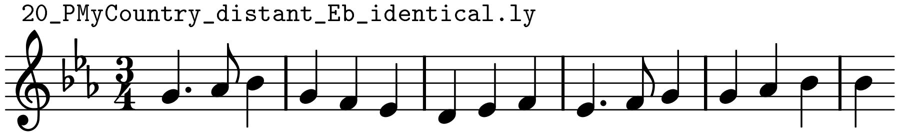
19_PMaryLamb_nearby_violate_outkey_F#.ly



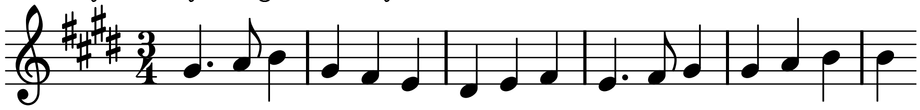
19_PMaryLamb_original_B.ly



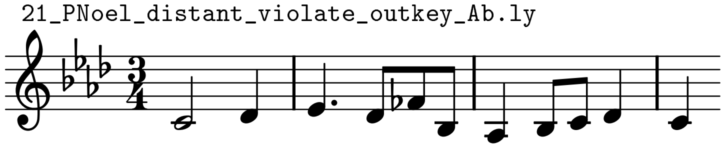
20_PMyCountry_distant_Eb_identical.ly



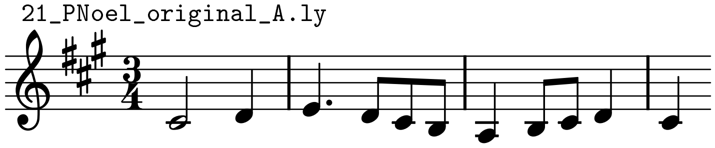
20_PMyCountry_original_E.ly



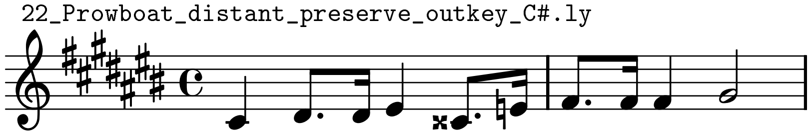
21_PNoel_distant_violate_outkey_Ab.ly



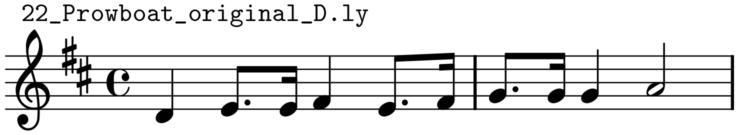
21_PNoel_original_A.ly



22_Prowboat_distant_preserve_outkey_C#.ly



22_Prowboat_original_D.ly

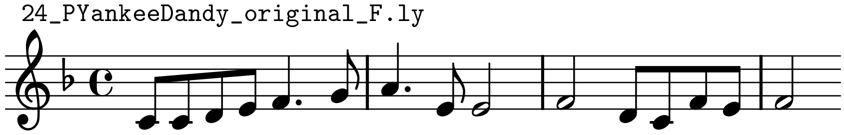


23_Pwewish_nearby_preserve_inkey_C.ly

23_Pwewish_original_G.ly

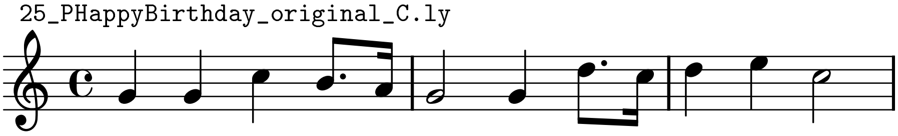
24_PYankeeDandy_distant_preserve_inkey_E.ly

24_PYankeeDandy_original_F.ly

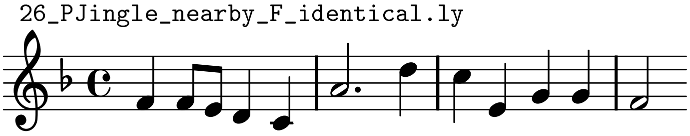


25_PHappyBirthday_nearby_G_identical.ly

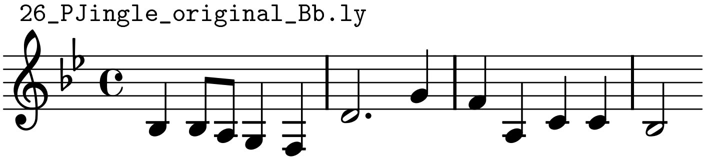
25_PHappyBirthday_original_C.ly



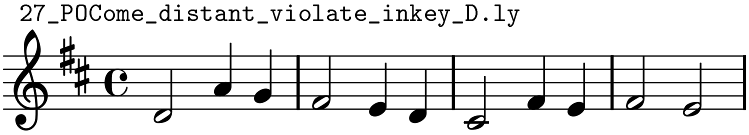
26_PJingle_nearby_F_identical.ly



26_PJingle_original_Bb.ly



27_POCome_distant_violate_inkey_D.ly



27_POCome_original_C#.ly

28_Poldmac_distant_preserve_outkey_A.ly

28_Poldmac_original_Ab.ly

E R functions

We have provided a scoring function as an R source file (gmsi_test_scoring.r). The method argument selects which scoring method you would like to run. If this argument is omitted, the function will select whether to run all the accuracy/discrimination analyses, or the classification analyses, based on the columns present in your datafile. The *Usage* sections of the test documentation describe the different scores for each test, and also which columns must be present for each kind of analysis. Finally, for carrying out classification analyses (e.g. for genre data), the function expects the binsize argument, which gives the size of each bin (or group) in the scoring solution (i.e. for 16 items sorted into 4 bins, this would be equal to 4).

The code below can be run by R novices; it shows a session in which a number of different data files are opened, scores obtained for each participant, and results are saved to disk. In this example, anything to the right of the character # is a comment and ignored by R until the end of the line. The function requires packages ROCR and e1071, which R should be able to install by itself if your computer is connected to the internet.

```
### get data from one or more participants...
### ...(see documentation for which columns to include)
mel.data <- read.csv(file="melody_data.csv", header=TRUE)
bat.data <- read.csv(file="bat_data.csv", header=TRUE)
gen.data <- read.csv(file="genre_data.csv", header=TRUE)

### load the supplied function...
### ...which is called 'performance.score'...
### ...e.g. from your current working directory...
source(file='gmsi_test_scoring.r')

### score melody data (use all methods) and save to file
mel.results <- performance.score(mel.data, method=c("acc", "dpr", "auc"))
write.csv(mel.results, file="melody_scores.csv")

### same for bat data
bat.results <- performance.score(bat.data, method=c("acc", "dpr", "auc"))
write.csv(bat.results, file="bat_scores.csv")

### something a bit different for genre data...
### ...using methods 'prs' and 'rnd'...
### ...and specifying binsize
gen.results <- performance.score(gen.data, method=c("prs", "rnd"), binsize=4)
write.csv(gen.results, file="genre_scores.csv")
```