View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Queen Mary Research Online

Modelling melodic discrimination tests: Descriptive and explanatory approaches

Peter M. C. Harrison, Jason Jiří Musil, and Daniel Müllensiefen

Goldsmiths, University of London

Author Note

Peter Harrison, Department of Psychology, Goldsmiths, University of London (corresponding author); Jason Musil, Department of Psychology, Goldsmiths, University of London; Daniel Müllensiefen, Department of Psychology, Goldsmiths, University of London.

Peter Harrison is now at the School of Electronic Engineering and Computer Science, Queen Mary, University of London.

The first author was supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). The second author was supported by a PhD bursary from the Psychology Department at Goldsmiths, University of London.

The authors would like to thank the reviewers for their very useful comments in the preparation of this manuscript. They are also grateful to Amit Avron for collecting the response data used in Study 2, which previously appeared in an unpublished Master's dissertation (Avron, 2012). The authors would also like to thank Bruno Gingras for producing some of the melodies used in v. 0.91 of the Gold-MSI melodic memory test.

Correspondence regarding this article should be addressed to Peter M. C. Harrison, School of Electronic Engineering and Computer Science, Queen Mary, University of London. Email: p.m.c.harrison@qmul.ac.uk.

Addresses

Peter Harrison, School of Electronic Engineering and Computer Science, Queen Mary, University of London, Mile End Road, London, E1 4NS. Email: p.m.c.harrison@qmul.ac.uk.

Jason Musil, Department of Psychology, Goldsmiths, University of London, Whitehead Building, New Cross, London, SE14 6NW. Email: j.musil@gold.ac.uk.

Daniel Müllensiefen, Department of Psychology, Goldsmiths, University of London, Whitehead Building, New Cross, London, SE14 6NW. Telephone: +44 (0)20 7919 7895. Email: d.mullensiefen@gold.ac.uk.

Abstract

Melodic discrimination tests have been used for many years to assess individual differences in musical abilities. These tests are usually analysed using classical test theory. However, classical test theory is not well-suited for optimising test efficiency or for investigating construct validity. This paper addresses this problem by applying modern item response modelling techniques to three melodic discrimination tests. First, descriptive item response modelling is used to develop a short melodic discrimination test from a larger item pool. The resulting test meets the test-theoretic assumptions of a Rasch item response model (Rasch, 1960) and possesses good concurrent and convergent validity as well as good testing efficiency. Second, an explicit cognitive model of melodic discrimination is used to generate hypotheses relating item difficulty to structural item features such as melodic complexity, similarity, and tonalness. These hypotheses are then tested on response data from three melodic discrimination tests (*n* = 317) using explanatory item response modelling. Results indicate that item difficulty is predicted by melodic complexity and melodic similarity, consistent with the proposed cognitive model. This provides useful evidence for construct validity. This paper therefore demonstrates the benefits of item response modelling both for efficient test construction and for test validity.

Keywords: melodic discrimination, similarity, memory, musical abilities, item response modelling

Modelling melodic discrimination tests: Descriptive and explanatory approaches

Melody is ubiquitous in the music of all cultures (e.g. Eerola, 2006; Schmuckler, 2009; Unyk, Trehub, Trainor, & Schellenberg, 1992). As a result, the ability to recognize, compare, and reproduce melodies is crucial both for the perception and the production of music. Correspondingly, melodic processing tests are commonly used to assess individual differences in musical aptitude and expertise (e.g. Bentley, 1966; Gaston, 1957; Gordon, 1965, 1982; Law & Zentner, 2012; Müllensiefen, Gingras, Musil, & Stewart, 2014; Seashore, 1919; Ullén, Mosing, Holm, Eriksson, & Madison, 2014; Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010; Wing, 1961).

Melodic processing abilities are typically assessed using melodic discrimination tests. In each trial of a melodic discrimination test, the test-taker is played several similar versions of an unfamiliar melody, and their task is to identify differences between these versions. The precise nature of the task can vary, but typically a ‘same-different’ task is used, where the test-taker has to determine whether two melody versions are the same or different (e.g. Law & Zentner, 2012; Müllensiefen et al., 2014; Wallentin et al., 2010). Sometimes the second melody is transposed in pitch relative to the first; in this case, the listener is instructed to ignore transposition and instead compare pitch intervals (e.g. Müllensiefen et al., 2014).

Melodic discrimination paradigms also form the basis of many melodic similarity experiments. As with many melodic discrimination tests, participants are typically presented with one pair of melodies in each trial. Instead of detecting differences between melodies, the participant’s task is to evaluate the similarity of these melodies (e.g. Eerola & Bregman, 2007; Müllensiefen & Frieler, 2007; Prince, 2014). However, whether the task is to evaluate melodic similarity or to detect melodic differences seems to make little difference to response patterns (Bartlett & Dowling, 1988). This suggests that similar cognitive processes underlie both scenarios.

Melodic discrimination tests are usually constructed and analysed using classical test theory (CTT; e.g. Gulliksen, 1950). The purpose of CTT is to model the statistical properties of scores delivered by ability tests. In CTT, analysis is primarily carried out at the level of the complete test, not the individual item. Each person is modelled as possessing a *true score* that would be achieved on the test if measurement error were zero; observed test scores are then produced by summing the true score together with an *error score* representing the test’s imprecision as a measurement instrument (e.g. Novick, 1966).

CTT has formed the basis of decades of test construction and validation. However, it possesses a number of important disadvantages, mostly stemming from its reliance on test-level analysis rather than item-level analysis. Firstly, CTT is not an ideal tool for choosing which items to include within a test. It provides some item-level measures of performance, such as mean scores and item-total correlations, but these performance measures are intrinsically confounded with one another, and they cannot be generalised to new tests or new test-taker populations (e.g. Schmidt & Embretson, 2003). This is problematic for efficient test construction, where it is important to ensure that each item contributes optimally to test performance. Secondly, CTT analyses at the test level can only provide limited information concerning construct validity, the question of how test scores relate to the underlying construct of interest (e.g. Messick, 1989, 1995). In these analyses, construct validity is primarily assessed by investigating external relationships between test scores and other measures. For example, evidence for construct validity comes when test scores correlate highly with other tests intended to measure the same or related abilities (concurrent and convergent validity) while correlating poorly with tests thought to measure unrelated abilities (discriminant validity). However, all of these arguments for construct validity depend themselves on the construct validity of the reference measures. In some senses, therefore, these arguments simply defer the problem of construct validity to other tests rather than addressing it directly.

The issue of item selection is problematic for melodic discrimination tests. Like many musical listening tests, melodic discrimination tests are often intrinsically inefficient because of the nature of the response paradigm. Each item usually only has a few response options, meaning that correct answers can often be achieved by guessing. This introduces noise into the response data, reducing test reliability. This can be compensated for with increased test length, but this comes at the expense of practicality and participant fatigue. In order to balance reliability with test length, it is therefore necessary to optimise item selection by choosing only the best performing items and ensuring that these items provide consistent discrimination power over the required ability range. Unfortunately, CTT is not well-suited to this task.

The issue of construct validity is also important for melodic discrimination testing. Despite the widespread use of melodic discrimination tests in musical listening test batteries, there is surprisingly little consensus about what cognitive ability (or abilities) these tests actually measure. Previous studies have proposed a range of underlying abilities, including ‘audiation’ (Gordon, 1989), melodic memory (Müllensiefen et al., 2014), and tonal memory (Vispoel, 1993). However, definitions of these abilities are usually cursory and unsubstantiated. This seriously undermines the construct validity of the melodic discrimination test.

This paper aims to address these issues of test efficiency and construct validity using modern techniques of *item response modelling* (also known as *item response theory*). Unlike the test-level focus of CTT, item response modelling is an approach to psychometric testing that focuses on analysing individual items. Two main approaches exist: descriptive modelling and explanatory modelling (de Boeck & Wilson, 2004). Descriptive modelling is a powerful tool for test construction, whereas explanatory modelling is a powerful tool for investigating construct validity.

Descriptive item response modelling uses response data to quantify the behaviour of each test item individually. Each item is treated as a black box, with the only feature of interest being its psychometric characteristics. Diagnostic checks can be used to assess the item’s psychometric quality, and information curves can be computed to illustrate how effective a particular item is for different ability levels. This information can then be used to select an optimal set of items for a future test.

Perhaps the most well-known example of descriptive item response modelling is the Rasch model (Rasch, 1960). In the Rasch model, each item is characterised by one difficulty parameter and each person by one ability parameter. The probability that a person responds correctly to a given item is modelled as a logistic function of the difference between person ability (β) and item difficulty (δ):

$$P(\text{success}) = \frac{e^{\beta-\delta}}{1+e^{\beta-\delta}}$$

When person ability is equal to item difficulty, the probability of success is 0.5; as person ability becomes much larger than item difficulty, the probability of success approaches 1, and so on. This model was later expanded to take account of further subtleties in response behaviour, such as differing item discrimination abilities and non-zero chance success rates (e.g. the three-parameter logistic model; Birnbaum, 1968; Lord, 1980). Nonetheless, the original Rasch model is still commonly used for test construction (e.g. Bond & Fox, 2015)

Descriptive item response modelling has been applied successfully once before to a melodic discrimination test (Vispoel, 1993). This item response model then formed the basis of a computerised adaptive test, where item selection was optimised on-the-fly according to the current performance of the test-taker. Computer simulations suggested that this should produce excellent improvements in testing efficiency. Unfortunately, this test never became widely available, and so there is still a demand for a short yet reliable melodic discrimination test.

Explanatory item response modelling provides an alternative approach to item response modelling where items are not treated as black boxes. Instead, explanatory item response models use structural features of items to explain their psychometric characteristics. Typically such a model will derive from an explicit cognitive model of the various mental processes involved in test-taking. By evaluating the fit of the item response model to the data, it is possible to test the cognitive model itself. Explanatory item response modelling therefore provides essential evidence for construct validity (Carroll, 1993; Embretson, 1983).

One of the first examples of explanatory item response models was the linear logistic test model (Fischer, 1973). This model extends the Rasch model by modelling item difficulty as a linear combination of structural item features, typically the number and type of fundamental cognitive operations required to answer the item correctly. Once a linear logistic test model is constructed, it is then possible to predict the difficulty of new items before they are administered to test-takers.

Though formal explanatory item response modelling has not yet been applied to the melodic discrimination paradigm, a great number of studies from the experimental psychology tradition have investigated how melodic discrimination performance is affected by item features (e.g. Cuddy, Cohen, & Mewhort, 1981; Cuddy, Cohen, & Miller, 1979; Cuddy & Lyons, 1981; Dowling & Bartlett, 1981; Dowling & Fujitani, 1971; Dowling, 1978; Mikumo, 1992; Schulze, Dowling, & Tillmann, 2012). However, these studies tend to focus solely on the role of working memory in melodic discrimination.

None of these studies have explicitly discussed the full range of cognitive processes inherent in the melodic discrimination task, or used their findings to substantiate the construct validity of the melodic discrimination test.

This paper therefore uses both descriptive and explanatory item response modelling to investigate the melodic discrimination paradigm. First, we review the various melodic discrimination tasks used in prior research. We then outline an explicit cognitive model for the task, conceptualising melodic discrimination as similarity comparison performed within the constraints of working memory. This model generates hypotheses relating item difficulty to structural item features, which are operationalised using formal measures of melodic complexity, similarity, and tonalness. Three empirical studies are then conducted. The first uses descriptive item response modelling to construct a short yet efficient melodic discrimination test. The second study assesses the construct validity of this melodic discrimination test by relating melodic discrimination scores to scores on other tests, thereby investigating concurrent and convergent validity. The final study then investigates construct validity by applying explanatory item response modelling to three different melodic discrimination tests. Through these complementary approaches, the aim is to address both the efficiency and the construct validity of the melodic discrimination test.

1. Melodic discrimination tasks

There are several different types of melodic discrimination task. Of these, the ‘same-different’ melodic discrimination task is probably the most common (e.g. Dowling & Fujitani, 1971). Here the participant is played two versions of the same melody which are either identical or non-identical after transposition. The participant is then asked to determine whether the two melodies are identical or not. In Gordon’s Advanced Measures of Music Audiation (AMMA; Gordon, 1989), the participant additionally has to state whether these melodies differ in pitch content or in rhythm content.

Another variant of the melodic discrimination paradigm requires the test-taker to identify which particular note differs between two versions of the same melody (Ullén et al., 2014; Vispoel, 1993). In some of these tests, every melody pair contains a difference somewhere (Ullén et al., 2014); in other tests, some melody pairs are allowed to be completely identical (Vispoel, 1993).

Some tests use more than two melodies per trial. Cuddy and colleagues (1979) played their participants one standard melody and two comparison melodies in each trial, and instructed them to determine which comparison melody matched the standard melody. Harrison (2015) played participants three melodies in each trial, and instructed them to identify which melody differed from the others.

We suggest that all of these task variants rely on very similar skills. On account of space constraints, this paper focuses on modelling the ‘same-different’ task. However, the model is expected to generalise well to other melodic discrimination tests.

2. Cognitive model

We propose that the essence of the melodic discrimination paradigm is a similarity comparison task that depends strongly on the limitations of working memory. In total, however, four important cognitive processes underlie the task: perceptual encoding, memory retention, similarity comparison, and decision-making. Though the final response is ultimately determined by the decision-making process, the reliability and accuracy of this decision depend on each of the preceding steps.

2.1. Perceptual encoding

Perceptual encoding applies to both melodies in the trial. In perceptual encoding, the listener forms a cognitive representation of a melody as it is played. This representation comprises a range of melodic features at various levels of abstraction, including pitch content, interval content, contour, tonality, and metrical structure. The difficulty of this task can vary depending on the nature of the melody. For example, some melodies exhibit a clearer harmonic structure than others, and presumably it is easier to derive a tonal representation for these melodies (e.g. Cuddy et al., 1981). Encoding difficulty may also depend on the prior musical context, typically the preceding melody in the trial. In particular, there is some evidence that tonal context may facilitate or impair the processing of the new melody, depending on the transposition between the two melodies (Cuddy et al., 1981, 1979; c.f. Takeuchi & Hulse, 1992).

2.2. Memory retention

Memory retention is only required for the first melody in each trial. The representation of this melody developed during perceptual encoding is stored in working memory so that it can eventually be compared to the second melody. However, because working memory is limited in capacity, the initial melody representation may not always be retained with complete precision. How well the melody is retained depends on the melody's memorability.

Complexity is an important contributor to memorability. More complex melodies are likely to place higher demands on the limited capacity of working memory, resulting in lower memorability. There are several different ways of operationalising melodic complexity; previous studies have used the number of notes in the melody (Akiva-Kabiri, Vecchi, Granot, Basso, & Schön, 2009; Brittin, 2000; DeWitt & Crowder, 1986; Edworthy, 1985; Schulze et al., 2012), and some studies have also used contour complexity (Croonen, 1994; Cuddy et al., 1981). While high length is reliably associated with poor melody discrimination performance, the effect of contour complexity seems less reliable.

Another contributing factor to memorability is the degree to which the melody conforms to culturally learned musical schemata, such as tonal and metrical structure. In general, stimuli which conform to learned schemata tend to be better retained in working memory than non-conforming stimuli (e.g. Egan & Schwartz, 1979; Engle & Bukstel, 1978; Gobet & Simon, 1998), and correspondingly melodies conforming to Western tonal structure tend to be easier to retain in working memory, at least for Western listeners (Cuddy et al., 1981; Dowling, 1991; Halpern, Bartlett, & Dowling, 1995; Schulze et al., 2012; Watkins, 1985). Similarly, metrical rhythmic patterns seem to be better retained in working memory than non-metrical patterns (Bharucha & Pryor, 1986).

2.3. Similarity comparison

In the similarity comparison process, the individual hears a new melody, compares it to the memory representation of a melody previously heard in that trial, and judges the similarity of the pair of melodies. In the standard 'same-different' task, this similarity judgement will be unidimensional, but in the AMMA variant the test-taker must make separate similarity judgments for pitch and rhythm dimensions. In both cases, we suggest that similarity judgments are made while the new melody is playing, meaning that this new melody does not need to be stored in working memory.

Melodic similarity is evaluated using the features available from the memory representation for the first melody. Of these features, tonality and contour seem to play the biggest role in determining similarity judgements, perhaps because these features dominate melodic working memory (Dowling, 1978; Schmuckler, 2009).

Several types of tonal similarity contribute to the similarity comparison process. One type concerns the key distance between the two melodies. Bartlett and Dowling (1980) found that pairs of melodies are perceived as more similar when the second melody is transposed to a related key (such as the dominant) as opposed to an unrelated key (such as the tritone). This key-distance effect seems to bias similarity judgments towards ‘same’ responses even when the listener is instructed to ignore transposition. Several studies have failed to replicate this effect, however (Pick et al., 1988; Takeuchi & Hulse, 1992).

A second type of tonal similarity concerns the melody’s harmonic implications after adjusting for transposition. Melodies that have different implications in terms of their underlying harmonic sequences will clearly be easier to distinguish. A simple example is when a note in a diatonic melody is substituted for a non-diatonic note (e.g. Cuddy et al., 1979).

Almost all experimental studies of melodic discrimination use stimuli where the comparison melodies are transposed within trials. This contrasts with musical listening test batteries, where it is common not to transpose the comparison melody (e.g. Gordon, 1989; Law & Zentner, 2012; Vispoel, 1993; Wallentin et al., 2010). The precise effects of this transposition are unclear. Dowling and Fujitani (1971) found melody discrimination to be much easier for untransposed melodies, perhaps because the similarity comparison process can make use of absolute pitch comparisons. Furthermore, the authors also found that contour similarity only played a role for transposed melodies, not untransposed melodies. However, this observed interaction between contour similarity and transposition may simply have been the artefact of a ceiling effect.

2.4. Decision-making

In the forced-choice version of the ‘same-different’ task, we suggest that the listener uses a certain similarity threshold as a decision criterion, and this threshold stays approximately constant throughout the test. Similarly to other perceptual threshold models in psychophysics (e.g. Gigerenzer & Murray, 1987), if perceived similarity exceeds this threshold, then the listener responds that the melodies are the same; otherwise, the melodies are deemed to be different. This assumption is implicit in studies that analyse the discrimination paradigm using signal detection theory (e.g. Müllensiefen et al., 2014; Schulze et al., 2012). In confidence-level versions of this paradigm, we assume instead that the listener’s stated confidence level corresponds directly to their similarity judgement (Bartlett & Dowling, 1988), which allows task performance to be assessed by calculating areas under the memory operating characteristic (e.g. DeWitt & Crowder, 1986; Dowling, 1971).

The discrimination task used in Gordon’s (1989) AMMA still uses two melodies in each trial, but the test-taker is given three response options: tonal difference, rhythmic difference, or no difference. There are several possible strategies the test-taker could employ here. We suggest one such strategy where the listener first decides whether or not a difference exists between the melodies, depending on whether the overall perceived similarity of the pair exceeds a certain threshold. If the threshold is exceeded, the participant responds ‘no difference’. Otherwise, the participant compares the pair’s rhythmic similarity to its pitch similarity. If the rhythmic similarity is lower, the participant responds ‘rhythmic difference’, otherwise the participant responds ‘tonal difference’.

2.5. Hypotheses

The cognitive model described above provides clear hypotheses about how item features should relate to item difficulty. Specifically, any item feature that impairs perceptual encoding, memory retention, similarity comparison, or decision-making should be expected to be positively associated with item difficulty. Differences in decision-making impairment are unlikely to arise within any one melodic discrimination test, since response paradigms typically do not change within a test. However, the remaining three stages are susceptible to effects of item features. Melodic complexity should impair memory retention, hence increasing item difficulty. Conformity to cultural schemata, such as tonal and metrical structure, should aid perceptual encoding and memory retention, decreasing item difficulty. Contour and tonal similarity should impair similarity comparison, hence increasing item difficulty. Transposition should impair perceptual encoding and similarity comparison, hence increasing item difficulty. Lastly, greater key distance between melodies should bias listeners towards responding ‘different’, hence decreasing difficulty for ‘different’ items and increasing difficulty for ‘same’ items.

The model also predicts that melodic discrimination performance should be affected by the order of the melodies being discriminated. This is particularly clear in the case of the ‘same-different’ task, where only the first melody in the pair needs to be retained in working memory. Suppose that the two melodies are different, and that one of these melodies is less memorable than the other. Since only the first melody needs to be retained in working memory, melodic discrimination performance should be worse when the less memorable melody comes first.

This asymmetry in melodic discrimination judgements has previously been documented by Bartlett & Dowling (1988). In this study, the authors presented participants with two melodies in each trial, one of which was scalar (i.e. comprised solely diatonic pitches; denoted *S*) and one of which was non-scalar (i.e. contained at least one non-diatonic pitch; denoted *N*). Scalar melodies should be more memorable than non-scalar melodies, since they conform better to Western musical schemata. Correspondingly, melodic discrimination performance should be better when the scalar melody comes first (*SN*) than when it comes second (*NS*). This is exactly what the authors found.

The same study from Bartlett & Dowling (1988) provides several additional results against which to test our model. Specifically, the best overall discrimination performance was found in *SN* trials; *SS* and *NS* trials both produced worse performance than *SN* trials, but approximately similar performance to each other; lastly, *NN* trials produced intermediate performance.

The first two results are clearly predicted by our model. *SN* trials benefit from both high memorability for the first melody presented (*S*) and low tonal similarity between *S* and *N* melodies, both of which are associated with good melodic discrimination performance. In contrast, *SS* trials possess high tonal similarity for the two melodies, resulting in worse performance than *SN* trials. Likewise, *NS* trials possess low memorability for the first melody (*N*) compared to *SN* trials, resulting in comparatively worse performance.

Our model does not make a clear prediction about the third result, the intermediate difficulty of *NN* trials. Though it is clear that memorability should be low in *NN* trials, it is not clear whether tonal similarity should be higher or lower for *NN* trials than in *NS* trials, as not all non-scalar melodies have equivalent harmonic implications. Nonetheless, the fact that *NN* trials elicited intermediate performance is consistent with our model, and suggests that tonal similarity was low for these melodies.

Interestingly, the original authors (Bartlett & Dowling, 1988) interpreted their results as a demonstration that memorability does not play a role in the asymmetry effect. Their rationale was that a memorability interpretation predicts that melodic discrimination performance should only be affected by the nature of the first melody in the pair. Therefore, performance should be just as high in *SS* trials as in *SN* trials, and performance should be just as bad in *NN* trials as in *NS* trials. When *SS* trials were in fact found to be harder than *SN* trials, the authors concluded that the memorability hypothesis had been contradicted.

As discussed above, however, their results can be easily explained as long as both memorability and similarity are taken into account. The effect of memorability explains why the task is harder when the first melody is non-scalar, but the effect of similarity explains why the task is harder when both melodies are scalar (*SS*) as opposed to when the second melody is non-scalar (*SN*). In conclusion, therefore, it seems that a wide range of experimental effects in melodic discrimination tasks can be explained by analysing the memorability of the first melody in the trial and the structural similarity of the pair of melodies in the trial.

3. Formal measures of melodic similarity, complexity, and tonalness

Structural item features need to be operationalised effectively if they are to form the basis of a reliable predictive model of item difficulty. Previous studies of melodic discrimination have manipulated melodic similarity, complexity, and tonalness as categorical variables. However, we suggest that these features may be better represented by continuous formal measures.

3.1. Melodic similarity

A great number of formal measures of melodic similarity already exist. These include geometric measures (Aloupis et al., 2006; O'Maidin, 1998), string-matching techniques such as edit distance (Crawford, Ilipoulos, & Raman, 1998; Mongeau & Sankoff, 1990), *n*-gram measures (Downie, 2003; Uitdenbogerd, 2002), hidden Markov models (Meek & Birmingham, 2001), and the Earth Mover's Distance algorithm (Typke, Wiering, & Veltkamp, 2007). There also exist measures derived directly from music theory (Grachten, Arcos, & de Mantaras, 2005) and from psychological models (Müllensiefen & Pendzich, 2009). Each of these classes of measures provides a useful perspective on melodic similarity, and it is difficult to choose just one and ignore the others.

One way to reconcile this diversity is by using hybrid measures, which combine scores across a number of different similarity measures to form one unidimensional measure of perceived similarity. An example is the hybrid measure *opti3* (Müllensiefen & Frieler, 2007), which was developed by modelling similarity judgements of pop songs by expert musicians. This measure takes a pair of melodies and outputs a numeric similarity rating between zero (completely dissimilar) and one (completely identical).

Previous research into the melodic discrimination task (e.g. Dowling, 1978; Schmuckler, 2009) suggests that similarity judgements in the paradigm rely primarily on contour and tonal similarity. Additionally, if the paradigm allows for rhythmic differences between melodies (e.g. Gordon, 1989), rhythmic similarity should presumably also predict item difficulty. We therefore construct a hybrid similarity measure out of individual measures of contour similarity, tonal similarity, and rhythmic similarity. These individual measures are sourced from the SIMILE toolbox (Müllensiefen & Frieler, 2004, 2007), where they are identified by the labels *conSEd*, *harmCorE*, and *rhytFuz2* respectively.

Each of these three measures works by computing new representations for each melody and then calculating edit distances between these representations. Contour representations are derived according to Steinbeck (1982). Essentially, this involves identifying all contour extrema, excluding those corresponding to changing notes, and then interpolating pitch values between these extrema using straight lines. Tonal representations take the form of sequences of harmonic symbols, one for each bar, each of which corresponds to the pitch class and mode of that bar as computed by the

Krumhansl-Schmuckler algorithm (Krumhansl, 1990). Lastly, the rhythmic representation is computed by classifying each note into one of five possible note-length classes: very short, short, normal, long, and very long. This classification is performed with respect to the notated beat length for the melody.

Once the respective representations are computed, similarity for a particular representation is calculated as the normalised edit distance between the two melody representations, as follows:

$$\sigma(s,t) = 1 - \frac{d_e(s,t)}{\max(|s|,|t|)}$$

where $\sigma(s,t)$ is the similarity between the two melody representations s and t, $d_e(s,t)$ is the edit distance between the melody representations s and t, and $|s|$ and $|t|$ correspond to the number of symbols in the melody representations s and t. Here a simple edit distance is used, meaning that the cost of inserting, deleting, or substituting a symbol is always one. Since the maximum edit distance between s and t is equal to the number of symbols in the longer of the two melody representations, and the minimum edit distance is zero, the similarity value ($\sigma(s,t)$) always takes a value between zero (completely different) and one (completely identical).

These three similarity measures are then combined linearly to form a unidimensional hybrid measure. Ideally the weights of each measure would be optimised empirically to match their relative perceptual contributions. However, because of a lack of prior empirical data, the present work uses equal weights for each measure.

3.2. Melodic complexity

Most previous studies have operationalised melodic complexity as the number of notes in the melody, and this measure has proved to be a reliable predictor of melodic discrimination difficulty (e.g. Akiva-Kabiri et al., 2009; DeWitt & Crowder, 1986; Schulze et al., 2012). In this paper we use the number of notes in the melody as well as two additional measures of melodic complexity. Both are calculated using the software toolbox FANTASTIC (Müllensiefen, 2009).

The first additional measure is *interval entropy*. Interval entropy describes how much intervallic variation there is within the melody. Let $F(i)$ denote the number of times that an interval of

i semitones occurs in the melody, with positive values of i denoting ascending intervals and negative values denoting descending intervals. Define the relative frequency of each interval as

$$f_i = \frac{F(i)}{\sum_j F(j)}$$

where j ranges over all intervals in the melody. Then interval entropy is defined as:

$$\text{interval entropy} = -\frac{\sum_i f_i \log_2 f_i}{\log_2 23}$$

Higher values of interval entropy correspond to greater intervallic variation.

The second additional measure is *step contour local variation*, which describes how much pitch varies at a local level. First, a step contour vector $\boldsymbol{x}$ is computed for the melody. This vector has length 64, and its elements correspond to samples of the raw pitch values (measured by MIDI note number) of the melody at equally spaced time intervals along the whole melody. Then step contour local variation is defined as the mean absolute difference between adjacent values in this vector:

$$\text{step contour local variation} = \frac{\sum_{i=1}^{63} |x_{i+1} - x_i|}{63}$$

These formal measures primarily address pitch complexity, not rhythmic complexity. The justification for this is that the melodic discrimination tests modelled in the present paper predominantly employ pitch differences between melodies, not rhythmic differences. As a result, we expect pitch memory to play a more important role than rhythmic memory in discrimination performance, and correspondingly pitch complexity should play a bigger role than rhythmic complexity in explaining item difficulty. However, for modelling tests where rhythmic differences play a big role, it would be worth including additional formal measures of rhythmic complexity.

3.3. Melodic tonalness

Conformity to Western tonal structure is assessed using the *tonalness* measure from the FANTASTIC toolbox (Müllensiefen, 2009), based on the Krumhansl-Schmuckler algorithm (Krumhansl, 1990). The total durations of each pitch class in the melody are correlated with the Krumhansl-Kessler (1982) profiles for all 24 major and minor keys, and tonalness is defined as the highest of the 24 correlation coefficients.

4. Studies

4.1. Study 1

The aim of the first study was to construct a short yet efficient melodic discrimination test. To do this, we develop a descriptive item response model of a longer pre-existing melodic discrimination test, and then use this model to select a set of items to maximise test performance for a typical population of adult students. This involves ensuring both that the retained items possess desirable psychometric characteristics individually, and also that these items combine to produce good discrimination power over an appropriate ability range.

Method

Participants

A total of 152 participants took part. These participants ranged from 18 to 39 years in age (*M* = 21.5, *SD* = 4.4), with approximately three quarters being female. All participants were first-year undergraduates who participated for course credit.

Materials

This study used the complete set of 28 melodic discrimination items from v. 0.91 of the melodic memory test component of the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014; Müllensiefen, Gingras, Stewart, & Musil, 2013). This test uses a 'same-different' discrimination paradigm where exactly half of the items constitute 'different' pairs. The second melody in each pair is always transposed relative to the first, either by a semitone or by a fifth. All melodies are between 10 and 17 notes in length, and were originally created by shuffling the order of intervals and rhythmic durations in pre-existing folk and popular melodies to render the original melodies unrecognisable. 'Different' pairs of melodies always differ in terms of the pitch of one, two, or three notes, and the nature of these differences is characterised by two systematically manipulated dichotomous variables: whether the difference violates the tonality of the original melody, and whether the difference violates the contour of the original melody.

Procedure

The Gold-MSI items were administered as part of a longer testing session collecting data for several unrelated studies. This testing session lasted approximately 100 minutes including short breaks between tests, and included two other listening tests, two questionnaires, and a visual attention and search test. Participants were tested in groups through a computerised interface, each with their own computer and headphones.

Results

In order to arrive at a smaller subset of the 28 items that would satisfy the rigorous assumptions of Rasch item response models (Rasch, 1960), we fit a Rasch model to the data from the 152 participants on all 28 items using the R package ‘eRM’ (Mair & Hatzinger, 2007). Andersen’s (1973) likelihood ratio test of the model assumption of subgroup homogeneity (median as split-criterion) on the resulting Rasch model failed ($p < .05$) and 8 items showed significant deviations ($p > .05$) from the model assumption according to subsequent Wald tests. After removing these 8 items, a second Rasch model was fitted which barely passed the likelihood ratio test at the conventional significance level ($p = .055$) but still contained 4 items that violated model assumptions. Next, individual items that failed the Wald test at a significance level of .05 were eliminated on a step-by-step basis, as is common procedure for reviewing and refining Rasch models (Bond & Fox, 2015). This iterative procedure arrived at a model passing the likelihood ratio test and containing 18 items that each passed the Wald test. Visually inspecting item difficulties and person abilities according to this model on a person-item map as well as the item information curves of all 18 items suggested that the test contained too many items at the easy end of the ability spectrum, and that several items had almost identical difficulty parameters and therefore seemed redundant. Therefore, the four easiest items as well as one item with a redundant difficulty parameter were removed and a Rasch model was fit to the remaining 13 items. This model passed the Andersen likelihood ratio test with all items also being clearly non-significant on the Wald test. This model also achieved a non-significant *p*-value ($p > .05$) on the goodness-of-fit chi-square test as implemented in the R package ‘ltm’ (Rizopoulos, 2006). In addition, the distribution of test scores (simple sum scores) on this 13-item test was balanced with respect to the low and high ends of the distribution and overall appeared fairly close to a normal distribution (a formal test of normality is not appropriate for the discrete distributions that arise from summing scores of 13 binary items).

Discussion

This study demonstrated the construction of a short 13-item melodic discrimination test which successfully met the requirements of the Rasch model. The resulting test takes about 6 minutes to complete, and is designed to possess good discrimination ability for an adult population of test-takers.

Rasch modelling has been extensively validated as a test construction tool (Bond & Fox, 2015), prompting its use in this study. However, there are two disadvantages to its use here. First, Rasch models assume a chance success rate of zero, whereas the 'same-different' task used in this test has a 50% chance success rate. Secondly, Rasch models cannot account for varying decision thresholds between participants. However, both of these problems can be mitigated by scoring the test with measures from signal detection theory and using sensitivity (*d'*) as the measure of test performance.

4.2. Study 2

The aim of this study was to investigate the construct validity of the shortened melodic discrimination test using measures of concurrent and convergent validity. Concurrent validity means that test scores correlate well with scores on a pre-established test of the same ability, whereas convergent validity means that test scores correlate appropriately with measures of other related abilities. Both are important indicators of construct validity.

Participants

Forty-four participants took part in this study, none of whom had participated in Study 1. These participants ranged in age from 18 to 63 years (*M* = 24.1, *SD* = 7.6), with exactly half being female and half male. All participants were students, and approximately a third received monetary remuneration in exchange for participation.

Materials

Melodic discrimination tests. Two melodic discrimination tests were used. The first was the new 13-item test constructed in Study 1. The second test was the Advanced Measures of Music Audiation (AMMA; Gordon, 1989), used to investigate concurrent validity. The AMMA constitute the most widely used melodic discrimination test in academic research, and feature in many recent studies as a measure of musical aptitude (e.g. Hu et al., 2013; Kühnis, Elmer, Meyer, & Jäncke, 2012, 2013; Mehr, Schachner, Katz, & Spelke, 2013; Mehr, Song, & Spelke, 2016). The AMMA comprise 30 items, none of which include transpositions. As described earlier, the AMMA use a variant of the 'same-different' task where the test-taker additionally has to identify whether alterations occur in pitch content or in rhythm content. Gordon recommends using responses to the melodies that differ in pitch content to calculate a tonal score for the participant, and those that differ in rhythm to calculate a rhythmic score. 'Same' items contribute to both scores.

Musical sophistication. Musical sophistication was assessed using the 39-item Gold-MSI questionnaire (Müllensiefen et al., 2014) in order to investigate convergent validity. This questionnaire assesses self-reported individual differences in skilled musical behaviours on five subscales (Active Musical Engagement, Perceptual Abilities, Musical Training, Singing Abilities, Emotional Engagement with Music) and one general factor (General Musical Sophistication).

Procedure

Data were collected as part of a larger validation study for the Gold-MSI questionnaire and listening tests (Avron, 2012). Participation was split into two testing sessions separated by 14 days. All testing was conducted in a quiet laboratory setting, with audio stimuli played over headphones. In the first testing session participants took the Weschler Abbreviated Scale of Intelligence (WASI; Wechsler, 2011; results not reported here), followed by the Gold-MSI questionnaire, and then the new melodic discrimination test. In the second session, participants first took the AMMA, then a repeat of the Gold-MSI questionnaire, then two tests of executive function and two unrelated Gold-MSI musical listening tests (results not reported here). Participants responded to the Gold-MSI questionnaire and the new melodic discrimination test over a computer interface, but responded to the AMMA using the official paper response sheet.

Results

Two participants failed to complete all tasks, but their remaining data are included in this analysis where possible. Sensitivity (*d'*) scores for the new melodic discrimination test were calculated using signal detection theory (Macmillan & Creelman, 2005). These scores were moderately correlated both with AMMA tonal scores (*r*(40) = .488, *p* = .001) and with AMMA rhythm scores (*r*(40) = .541, *p* < .001). Tonal and rhythm scores from the AMMA also correlated very highly with each other (*r*(41) = .825, *p* < .001).

Self-report scores from the Gold-MSI questionnaire were averaged between the two testing sessions before being compared to scores on the new melodic discrimination test. Melodic discrimination *d*' scores were significantly correlated with General Musical Sophistication (*r*(41) = .412, *p* = .006) as well as with Active Musical Engagement (*r*(41) = .419, *p* = .005), Perceptual Abilities (*r*(41) = .436, *p* = .003) and Singing Abilities (*r*(41) = .333, *p* = .029). However, test performance was not significantly correlated with Musical Training (*r*(41) = .246, *p* = .112) or with Emotional Engagement with Music (*r*(41) = .217, *p* = .162).

Discussion

The new melodic discrimination test demonstrated good concurrent validity as evidenced by moderate correlations with Gordon's AMMA and good convergent validity as evidenced by correlations with several self-reported measures of musical sophistication. The high correlation between the tonal and rhythm scores of the AMMA suggests that they both assess shared abilities, although part of this correlation will come from the fact that 'same' items contribute to both scores.

The lack of significance of the correlation between test scores and musical training was surprising but may have been an artefact of the small sample group used. A study of the original (v. 0.91) Gold-MSI melodic memory test observed a correlation of *r* = .301 between melodic discrimination scores and musical training, and this correlation was highly statistically significant on account of the study’s large sample size (*N* = ~140,000; Müllensiefen et al., 2014). However, a correlation of *r* = .301 would only have approximately a 50% chance of producing a statistically significant effect with only 43 participants, as in the present study. The relationship between melodic discrimination ability and musical training is therefore re-examined in Study 3.

4.3. Study 3

The aim of this study was to investigate the construct validity of the melodic discrimination paradigm using explanatory item response modelling. One short test does not provide enough variation in item features to explore their effects on item parameters properly, so this study compiled response data for three different melodic discrimination tests: v. 0.91 of the Gold-MSI melodic memory test (reanalysing data from Study 1), the AMMA (reanalysing data from Study 2), and the Musical Ear Test (MET; Wallentin et al., 2010; data collected in this study).

Construct validity can be supported by explanatory item response modelling to the extent that variations in item difficulty can be predicted from a cognitive understanding of the task involved (Embretson, 1983). The cognitive model presented in this paper provides clear predictions about how item features should predict item difficulty. Specifically, melodic complexity and similarity should increase item difficulty, tonalness should decrease difficulty, and transposition should increase difficulty. This provides four experimental hypotheses to be tested by this study.

Contrary to prior research, Study 2 had found that musical training did not predict melodic discrimination performance. An additional aim of Study 3 was therefore to reinvestigate the possible association between musical training and task performance.

Method

Participants

This study used data from 317 participants. Of these, 156 participants came from Study 1, with four extra participants contributing data after the test construction process described in Study 1 was completed. Study 2 provided data from 42 participants. An additional 119 participants were then recruited for Study 3. This participant group was recruited by a market research company[1] and was nationally representative in terms of age, gender, occupation, income, and geographic location. Participant ages ranged from 18 to 77 (*M* = 42.6, *SD* = 14.4), and approximately half of the participants were female.

Materials

The materials used in Studies 1 and 2 have already been described. Study 3 additionally made use of the MET (Wallentin et al., 2010), a listening battery containing a 52-item melodic discrimination test using the 'same-different' melodic discrimination paradigm. Like the AMMA, this test does not contain any transpositions. In the present study we use only the first 20 items of the melodic subtest, thereby shortening its length to approximately 4 minutes. Using the Spearman-Brown prophecy formula, it was calculated that this shortened test would still possess good internal reliability (estimated Cronbach's α = 0.90). All melodies are between 3 to 8 notes in length and have a duration of one bar. The 'different' trials all contain one pitch violation, and in half of these cases this pitch violation also constitutes a contour violation.

Procedure

The procedures for Studies 1 and 2 have already been described. Data for Study 3 were collected as part of a validation study for a series of computerised adaptive listening tests (Harrison, 2015). This validation study took place online using the Concerto testing platform (Scalise & Allen, 2015). Participants took the MET at the end of a 30-minute testing session comprising two other listening tests and a short questionnaire. Participants agreed to wear headphones and to take the experiment in a quiet room free from interruptions.

Item analysis

The 20 MET items and the 30 AMMA items were transcribed manually from the original audio files, and along with the 28 Gold-MSI items were converted to tabular format with numerical

[1] http://www.qualtrics.com

values characterising note pitch and onsets. Formal measures of melodic complexity, similarity, and tonalness were then calculated according to the definitions provided earlier.

Each of the three melodic discrimination tests used in this study involves just one melody comparison per trial. According to the cognitive model proposed in this paper, it is only the first melody in a pair that needs to be maintained in memory, and therefore only this melody should be considered when assessing complexity and tonalness. On this basis, each item's measures of complexity and tonalness were calculated solely from the first melody in the pair.

Results

Comparing item features for the three tests

Pairwise correlation coefficients were calculated for all the item features. The results indicated that length (i.e. number of notes), interval entropy, and step contour local variation were all strongly positively correlated (for each pair, $r(76) > .50$, $p < .001$). No other item features were significantly correlated.

On account of their high collinearity, the three features length, interval entropy, and step contour local variation were combined using principal component analysis to form a composite measure of melodic complexity. Different sets of weightings for this composite measure were estimated for 'same' items and for 'different' items, but in both cases, all three variables loaded approximately equally onto these composite measures (loadings: $.80 < x < .90$).

Distributions of melodic complexity, similarity, and tonalness are plotted in Figures 1–3. Though the three tests differ systematically on these measures, there is on the whole substantial overlap between scores on different tests, and there is significant variation in scores within each test, which bodes well for explanatory item response modelling. One possible exception is tonalness, where values are universally high, reflecting the fact that the great majority of melodies in this corpus were tonal.

All items in the AMMA and the MET use untransposed melodies, whereas all the items in the Gold-MSI use transposed melodies. Unfortunately, this meant that effects of transposition on item difficulty would be confounded by differences in abilities between participant groups used for the different studies. Transposition was therefore dropped from the analysis.

Designing the explanatory item response models

We construct our explanatory item response models within the framework of generalised linear mixed modelling (de Boeck et al., 2011; Doran, Bates, Bliese, & Dowling, 2007). Mixed-effects logistic regression can be used to construct an explanatory version of the Rasch model used in Study 1, where item difficulty is instead modelled as a linear combination of predictor variables. However, the standard Rasch model is not generally well-suited to modelling melodic discrimination tasks because it assumes a zero chance success rate, whereas most melodic discrimination tasks in fact have relatively high chance success rates.

In order to account for these non-zero chance success rates, we modify the link function (γ) within the logistic regression to produce a non-zero lower asymptote in the response function, as follows:

$$\gamma(x) = \log\left(\frac{x-c}{1-x}\right)$$

where c corresponds to the probability (between 0 and 1) of guessing the answer correctly by chance (the *guessing parameter*), and *x* corresponds to the expected success rate.

Participants and items are specified as random effects and proposed predictors of item difficulty are specified as fixed effects. By extracting the coefficients of the fixed effects, a linear model can be constructed that predicts item difficulty on the basis of the proposed predictors. The resulting explanatory item response model is analogous to a three-parameter logistic model (Birnbaum, 1968), but with a pre-specified guessing parameter. This *a priori* constraint is useful as the empirical estimation of guessing parameters typically requires a great number of participants (e.g. de Ayala, 2009)

An additional advantage of constructing explanatory item responses models using mixed-effects modelling is the ability to account for hierarchical characteristics of the response data (e.g. Doran et al., 2007). This ability is crucial for modelling the current dataset, where data are aggregated from three different studies. These three different studies differ systematically in many ways: The participants were sampled from different populations; some participants took their tests online whereas some took their tests in the lab; some tests use transposition and some do not; the tests use different timbres and tempi; and so on. It is important to take these differences into account when aggregating the three datasets. To do this we make the following assumptions. First, we acknowledge that the three different participant groups may differ systematically in terms of ability. We therefore model participant ability within each group as being sampled from separate normal distributions each with different means and variances. Secondly, we assume that the differences in implementations between the tests, such as the transposition of the melodies, the user interface, and the motivation of the participants, can all be modelled as a numeric constant for each test that is added or subtracted to the item difficulty of each item within that test. This constant can be described as the test's *inherent difficulty*. After these differences between the tests are taken into account, it is then understood that part of the remaining variation in item difficulties can be accounted for as the result of systematic effects of structural item features: complexity, similarity, and tonalness. Any remaining variation should then be uncorrelated error.

We fit these item response models using the 'lme4' and 'psyphy' packages (Bates, Maechler, Bolker, & Walker, 2015; Knoblauch, 2014) implemented in the statistical computing software 'R' (R Core Team, 2014). The mean ability for each test's participant group is combined with each test's inherent difficulty to form one three-level categorical variable, *test*, which is estimated as a fixed effect. Abilities of individual participants are estimated as random intercepts, with separate variance parameters estimated for each participant group. Hypothesised effects of structural item features are modelled as fixed effects. Residual variation in item difficulty is modelled as a random intercept for each item. Except for where otherwise stated, this specification of the item response model is used for all further analyses.

One limitation of this modelling approach is that the same guessing parameter must be employed for all items. This is problematic for the current dataset, as the AMMA have three response options (i.e. guessing parameter of 0.33) compared to the two options of the Gold-MSI and MET (i.e. guessing parameter of 0.5). In these analyses we adopt a guessing parameter of 0.33 for all items. The rationale behind using the lower guessing parameter is that, owing to the bounded property of the logistic function, underestimating the guessing parameter results in better model fit than overestimating it.

We model ‘same’ and ‘different’ items separately because our cognitive model of performance on ‘same-different’ melodic discrimination tasks would predict that the item difficulty predictors should behave differently for these two types of items. In particular, melodic similarity is hypothesised to be positively correlated with item difficulty for ‘different’ items, whereas the corresponding relationship is meaningless for ‘same’ items because all ‘same’ items have perfect similarities by definition. Modelling ‘same’ and ‘different’ items separately is a simple way to avoid this problem. Usefully, this also eliminates the issue of participant bias that the Rasch model was unable to account for, as separate participant intercepts are estimated for ‘same’ and for ‘different’ items.

Results of explanatory item response modelling

‘Same’ items

A generalised linear mixed model was constructed for the ‘same’ items according to the procedures described above, with all continuous predictors scaled and centred to give standard deviations of 1 and means of 0. The null model (Model 0) comprised a fixed effect of test, a random intercept for items, and three random intercepts for participants, one for each test. Model 1 was constructed by taking the null model and adding a fixed effect of musical training. Model 2 was then constructed by taking Model 1 and adding a fixed effect of complexity. Lastly, Model 3 was constructed from Model 2 by adding a fixed effect of tonalness. The resulting four models are compared in Table 1. The likelihood ratio tests indicate that the addition of musical training and complexity significantly improved model fit, whereas the addition of tonalness did not. These results are supported by the differences in Akaike’s information criterion (AIC) values, which indicate very strong support for Model 1, moderate support for Model 2, and no support for Model 3 (Murtaugh, 2014). However, the more conservative Bayesian information criterion (BIC) analysis supports the addition of musical training but not the addition of complexity or tonalness.

The coefficients for the fixed effects can be converted to a traditional item response theory metric (e.g. de Ayala, 2009; DeMars, 2010) by dividing by the standard deviation of the random effect of participant. Here we use the random effect for the MET participants, since this participant group was most representative of the general population. Taking Model 2 as the final model, the fixed-effect coefficients indicate that increasing musical training by one standard deviation increased ability by 0.26 (*SE* = 0.08, *p* < .001), while increasing complexity by one standard deviation increased item difficulty by 0.63 (*SE* = 0.22, *p* = .004).

The residual variation in item difficulty can similarly be estimated by taking the standard deviation of the item random intercept and dividing it by the standard deviation of the participant random intercept. This estimates the error standard deviation in item difficulty to be 0.92.

‘Different’ items

Data for the ‘different’ items were modelled in a similar manner, with the continuous predictors scaled and centred. The null model (Model 0) took the same form as for the ‘same’ items, and then four models were constructed for comparison by sequentially adding musical training (Model 1), complexity (Model 2), melodic similarity (Model 3), and tonalness (Model 4) as fixed effects. Model 4 did not converge, probably because of the complexity of the model, but the four simpler models did converge, and are compared in Table 2. The likelihood ratio tests indicate that musical training, complexity, and melodic similarity each significantly improved the fit of the model. The AIC values show substantial support for musical training and minor support for complexity and melodic similarity. The BIC values, meanwhile, only show support for musical training. The fixed-effect coefficients in Model 3 indicate that increasing musical training by one standard deviation increased ability by 0.82 (*SE* = 0.11, *p* < .001), increasing melodic similarity by one standard deviation increased item difficulty by 0.50 (*SE* = 0.20, *p* = .013), and increasing complexity by one standard deviation gave a marginally significant increase of 0.57 in item difficulty (*SE* = 0.28, *p* = .064). The estimated error standard deviation in item difficulty was 0.90.

Discussion

The aim of this final study was to investigate the construct validity of the melodic discrimination paradigm using explanatory item response modelling. A cognitive model for the melodic discrimination task was used to generate testable hypotheses relating structural item features to item difficulty. On the basis of this cognitive model, it was hypothesised that melodic complexity and similarity should increase item difficulty, tonalness should decrease difficulty, and transposition should increase difficulty.

The data provide positive support for two of these hypotheses. Both melodic complexity and similarity were positively related to item difficulty, as predicted. The predictive power of these predictors was supported by likelihood ratio tests and AIC statistics, but not by BIC statistics. However, BIC statistics are generally only appropriate when the exact ‘true’ model is contained in the candidate set of models, something which is very unlikely in cognitive experiments such as this (Murtaugh, 2014). It seems reasonable, therefore, to accept the support of the AIC statistics.

Unfortunately, it was not possible to investigate the effect of transposition properly, because two out of the three melodic discrimination tests did not contain any transposed melodies. A similar effect may have prevented tonalness from having an effect on item difficulty, since while the MET employs some atonal melodies, both the AMMA and the Gold-MSI use only tonal melodies. The effects of both transposition and tonalness will therefore have largely been subsumed by the fixed effect of *test*.

It was also hypothesised that musical training should be associated with better performance on the melodic discrimination task. The data strongly support this hypothesis, matching previous research linking melodic discrimination ability to musical expertise (e.g. Dowling, Bartlett, Halpern, & Andrews, 2008; Müllensiefen et al., 2014; Wallentin et al., 2010).

The high collinearity between the three complexity measures was to be expected. Melodies with more notes correspondingly possess more intervals, and this provides more opportunity for variety in the intervallic distribution, resulting in a correlation between length and interval entropy. Likewise, a melody with more notes has more opportunity for pitch variation, producing a correlation between length and step contour local variation. An informal analysis we conducted of a corpus of Irish folk melodies confirmed that these high correlations are not an artefact of the present corpus.

Despite their high correlations, length, interval entropy, and step contour local variation each should capture some unique facet of complexity. Unfortunately, because we were using pre-existing melodic discrimination tests, it was not possible to manipulate these three features orthogonally, and so their relative combinations to item difficulty could not be evaluated. This could provide the basis for an interesting follow-up experimental study.

The unexplained variation in item difficulty was rather high for both the ‘same’ and the ‘different’ items. This suggests that there is still substantial room for improvement for the explanatory model. This might be achieved by using improved formal measures of melodic complexity and similarity, for example by optimising the weights of the hybrid similarity measure or by using a weighted edit distance rather than a simple edit distance for the similarity measures. The item response model might also be improved by developing the cognitive model further, and using it to identify additional predictors of item difficulty.

5. General discussion

The aim of this paper was to address the efficiency and validity of traditional melodic discrimination tests using modern techniques of item response modelling. Two complementary approaches were used: a descriptive approach and an explanatory approach.

Studies 1 and 2 used descriptive item response modelling to construct and validate a new short yet efficient melodic discrimination test. This test satisfies the assumptions of a Rasch model, shows good concurrent and convergent validity, and is freely available for research[2]. However, as a trade-off for its short duration, the test’s discriminative power is necessarily limited. Moreover, because it is calibrated on a sample of a general student population, it will have low discrimination power in population groups with very high ability (e.g. professional musicians) or very low ability (e.g. amusics).

Study 3 used explanatory item response modelling to investigate the construct validity of the melodic discrimination test. On the basis of a cognitive model of melodic discrimination, hypotheses

[2] http://www.gold.ac.uk/music-mind-brain/gold-msi/

were generated relating item difficulty to structural item features. These hypotheses were then tested using response data from pre-existing melodic discrimination tests. The results support the proposed cognitive model, making an important contribution to the construct validity of the melodic discrimination test.

This paper demonstrates that the melodic discrimination task cannot automatically be characterised as a simple measure of a single cognitive ability, such as melodic memory. Instead, melodic discrimination must draw on a number of distinct cognitive processes, each of which may contribute to individual differences in melodic discrimination ability. It is important to take this into account when interpreting scores on melodic discrimination tests, instead of simply equating test scores with concepts such as musical aptitude (e.g. Hu et al., 2013; Mehr et al., 2013, 2016) or melodic memory (e.g. Müllensiefen et al., 2014; Zenatti, 1975).

It should be acknowledged, however, that the fact that melodic discrimination relies on several cognitive abilities does not necessarily mean that each ability contributes equally to individual differences in melodic discrimination scores. For example, it could be the case that almost all individuals possess sufficient perceptual encoding abilities for the traditional melodic discrimination task, meaning that perceptual encoding ability never is a limiting factor in task performance. Another alternative is that perceptual encoding is indeed a limiting factor in task performance, but individual variation in perceptual encoding abilities is small, and so this variation does not contribute substantially to individual differences in melodic discrimination scores. The methodologies used in this paper were not well-suited to investigating these questions. However, future research could aim to separate these different abilities through the use of multidimensional latent trait models, or by combining response data from a greater variety of testing paradigms, such as melodic recall tasks (e.g. Boltz, 1991) and similarity judgement tasks (e.g. Müllensiefen & Frieler, 2007).

In this paper we used descriptive item response modelling for test construction and explanatory modelling to investigate construct validity. However, explanatory item response modelling may also prove to be an exciting tool for test construction. Effective explanatory item response models can be used to predict item parameters for computerised adaptive tests, drastically reducing their production costs (e.g. Gierl, 2013). Such computerised adaptive tests can be remarkably efficient, requiring many fewer items to match the reliability of equivalent non-adaptive tests (de Ayala, 2009; Linden & Glas, 2007), and maintaining a high discriminative power regardless of the ability level being tested. As a result, developing good explanatory item response models for melodic discrimination could enable the economical construction of much more efficient melodic discrimination tests. We hope that the present paper provides a useful step towards this goal.

References

Akiva-Kabiri, L., Vecchi, T., Granot, R., Basso, D., & Schön, D. (2009). Memory for tonal pitches: A music-length effect hypothesis. *Annals of the New York Academy of Sciences*, *1169*, 266–269. doi:10.1111/j.1749-6632.2009.04787.x

Aloupis, G., Fevens, T., Langerman, S., Matsui, T., Mesa, A., Nuñez, Y., … Toussaint, G. (2006). Algorithms for Computing Geometric Measures of Melodic Similarity. *Computer Music Journal*, *30*(3), 67–76. doi:10.1162/comj.2006.30.3.67

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140.

Avron, A. (2012). *Reliability and validity of the Gold-MSI, and links between musicality and intelligence*. Master's dissertation, Goldsmiths College, University of London.

Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(3), 501–515.

Bartlett, J. C., & Dowling, W. J. (1988). Scale structure and similarity of melodies. *Music Perception*, *5*(3), 285–314.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Linear mixed-effects models using Eigen and S4. Retrieved from https://cran.r-project.org/package=lme4

Bentley, A. (1966). *Measures of musical abilities*. London, England: George A. Harrap.

Bharucha, J. J., & Pryor, J. H. (1986). Disrupting the isochrony underlying rhythm: An asymmetry in discrimination. *Perception & Psychophysics*, *40*(3), 137–141. doi:10.3758/bf03203008

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Boltz, M. (1991). Some structural determinants of melody recall. *Memory & Cognition*, *19*(3), 239–251. doi:10.3758/BF03211148

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Brittin, R. V. (2000). Perception and recall of melodic and rhythmic patterns: Effects of example length and tonal/rhythmic variety. In C. Mizener (Ed.), *Texas music education research 2000* (pp. 17–25). Austin, TX: Texas Music Educators Association.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Crawford, T., Ilipoulos, C. S., & Raman, R. (1998). String-matching techniques for musical similarity and melodic recognition. In W. B. Hewlett & E. Selfridge-Field (Eds.), *Computing in musicology: Vol. 11. Melodic similarity: Concepts, procedures, and applications* (pp. 73–199). Cambridge, MA: MIT Press.

Croonen, W. L. (1994). Effects of length, tonal structure, and contour in the recognition of tone series. *Perception & Psychophysics*, *55*(6), 623–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8058450

Cuddy, L. L., Cohen, A. J., & Mewhort, D. J. K. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 869–883.

Cuddy, L. L., Cohen, A. J., & Miller, J. (1979). Melody recognition: The experimental application of musical rules. *Canadian Journal of Psychology*, *33*(3), 148–157. doi:10.1037/h0081713

Cuddy, L. L., & Lyons, H. I. (1981). Musical pattern recognition: A comparison of listening to and studying tonal structures and tonal ambiguities. *Psychomusicology: A Journal of Research in Music Cognition*, *1*(2), 15–33. doi:10.1037/h0094283

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

de Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011).

The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12).

de Boeck, P., & Wilson, M. (2004). Descriptive and explanatory response models. In *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). New York, NY: Springer. doi:10.1007/978-1-4757-3990-9

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

DeWitt, L. A., & Crowder, R. G. (1986). Recognition of novel melodies after brief delays. *Music Perception*, *3*(3), 259–274.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2), 1–18. doi:10.1111/j.1467-9868.2007.00600.x

Dowling, W. J. (1971). Recognition of inversions of melodies and melodic contours. *Perception & Psychophysics*, *9*(3), 348–349. doi:10.3758/BF03212663

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*(4), 341–354.

Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception & Psychophysics*, *50*(4), 305–313.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology: A Journal of Research in Music Cognition*, *1*(1), 30–49.

Dowling, W. J., Bartlett, J. C., Halpern, A. R., & Andrews, M. W. (2008). Melody recognition at fast and slow tempos: Effects of age, experience, and familiarity. *Perception & Psychophysics*, *70*(3), 496–502. doi:10.3758/PP

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *The Journal of the Acoustical Society of America*, *49*(2B), 524–531. doi:10.1121/1.1912382

Downie, J. S. (2003). Music Information Retrieval. In B. Cronin (Ed.), *Annual Review of Information Science and Technology 37* (pp. 295–340). Medford, NJ: Information Today. doi:10.1002/aris.1440370108

Edworthy, J. (1985). Interval and contour in melody processing. *Music Perception*, *2*(3), 375–388.

Eerola, T. (2006). Perceived complexity of western and African folk melodies by western and African listeners. *Psychology of Music*, *34*(3), 337–371. doi:10.1177/0305735606064842

Eerola, T., & Bregman, M. (2007). Melodic and contextual similarity of folk song phrases. *Musicae Scientiae, Disc.4A*, 211–233. doi:10.1177/102986490701100109

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*(2), 149–58. doi:10.3758/BF03197595

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197. doi:10.1037/0033-2909.93.1.179

Engle, R. W., & Bukstel, L. H. (1978). Memory processes among bridge players of differing expertise. *The American Journal of Psychology*, *91*(4), 673–689. doi:http://dx.doi.org/10.2307/1421515

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. doi:10.1016/0001-6918(73)90003-6

Gaston, E. T. (1957). *A test of musicality: Manual of Directions*. Lawrence, KA: Odell's Instrumental Service.

Gierl, M. J. (2013). Automatic item generation: An introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: revisiting the chunking hypothesis. *Memory, 6*(3), 225–255. doi:10.1080/741942359

Gordon, E. E. (1965). *Musical aptitude profile*. Boston, MA: Houghton Mifflin.

Gordon, E. E. (1982). *Intermediate measures of music audiation*. Chicago, IL: G.I.A. Publications.

Gordon, E. E. (1989). *Advanced measures of music audiation*. Chicago, IL: G.I.A. Publications.

Grachten, M., Arcos, J. L., & de Mantaras, R. L. (2005). Melody retrieval using the Implication/Realization model. *MIREX-ISMIR 2005: 6th International Conference on Music Information Retrieval, London 2005*.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

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.

Harrison, P. M. C. (2015). *Constructing computerised adaptive tests of musical listening abilities*. Master’s dissertation, Goldsmiths College, University of London.

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain and Language, 127*(3), 366–376. doi:10.1016/j.bandl.2012.11.006

Knoblauch, K. (2014). psyphy: Functions for analyzing psychophysical data in R. Retrieved from http://cran.r-project.org/package=psyphy

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. New York, NY: Oxford University Press.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368. doi:10.1037/0033-295X.89.4.334

Kühnis, J., Elmer, S., Meyer, M., & Jäncke, L. (2012). Musicianship boosts perceptual learning of pseudoword-chimeras: An electrophysiological approach. *Brain Topography*, *26*(1), 110–125. doi:10.1007/s10548-012-0237-y

Kühnis, J., Elmer, S., Meyer, M., & Jäncke, L. (2013). The encoding of vowels and temporal speech cues in the auditory cortex of professional musicians: An EEG study. *Neuropsychologia*, *51*(8), 1608–1618. doi:10.1016/j.neuropsychologia.2013.04.007

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PLoS ONE*, *7*(12), e52508. doi:10.1371/journal.pone.0052508

Linden, W. J. van der, & Glas, C. A. W. (2007). Statistical aspects of adaptive testing. *Handbook of Statistics*, *26*(6), 801–838. doi:10.1016/S0169-7161(06)26025-5

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. New York, NY: Lawrence Erlbaum.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal Of Statistical Software*, *20*(9), 1–20.

Meek, C., & Birmingham, W. P. (2001). Thematic extractor. In *Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (ISMIR 2001)* (pp. 119–128).

Mehr, S. A., Schachner, A., Katz, R. C., & Spelke, E. S. (2013). Two randomized trials provide no consistent evidence for nonmusical cognitive benefits of brief preschool music enrichment. *PLoS ONE*, *8*(12). doi:10.1371/journal.pone.0082007

Mehr, S. A., Song, L. A., & Spelke, E. S. (2016). For 5-month-old infants, melodies are social. *Psychological Science*. doi:10.1177/0956797615626691

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741

Mikumo, M. (1992). Encoding strategies for tonal and atonal melodies. *Music Perception*, *10*(1), 73–82.

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*(3), 161–175.

Müllensiefen, D. (2009). FANTASTIC: Feature ANalysis Technology Accessing STatistics (In a Corpus): Technical Report v. 1.5. London: Goldsmiths, University of London. Retrieved from http://www.doc.gold.ac.uk/isms/m4s/FANTASTIC_docs.pdf

Müllensiefen, D., & Frieler, K. (2004). Melodic Similarity: Approaches and applications. In *Proceedings of the 8th International Conference on Music Perception & Cognition* (pp. 283–289).

Müllensiefen, D., & Frieler, K. (2007). Modelling experts' notions of melodic similarity. *Musicae Scientiae*, *Disc.4A*, 183–210. doi:10.1177/102986490701100108

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, *9*(2). doi:10.1371/journal.pone.0089642

Müllensiefen, D., Gingras, B., Stewart, L., & Musil, J. (2013). Goldsmiths Musical Sophistication Index (Gold-MSI) v1.0: Technical Report and Documentation Revision 0.3. Goldsmiths, University of London. Retrieved from http://www.gold.ac.uk/music-mind-brain/gold-msi/publications/

Müllensiefen, D., & Pendzich, M. (2009). Court decisions on music plagiarism and the predictive value of similarity algorithms. *Musicae Scientiae*, *13*(1 Suppl), 257–295. doi:10.1177/102986490901300111

Murtaugh, P. A. (2014). In defense of P values. *Ecology*, *95*(3), 611–617.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of*

Mathematical Psychology, *3*(1), 1–18. doi:10.1016/0022-2496(66)90002-2

O’Maidin, D. (1998). A geometrical algorithm for melodic difference in melodic similarity. In W. B. Hewlett & E. Selfridge-Field (Eds.), *Melodic similarity: Concepts, procedures, and applications (Computing in Musicology 11)*. Cambridge, MA: The MIT Press.

Pick, A. D., Palmer, C. F., Hennessy, B. L., Unze, M. G., Jones, R. K., & Richardson, R. M. (1988). Children’s perception of certain musical properties: Scale and contour. *Journal of Experimental Child Psychology*, *45*, 28–51.

Prince, J. B. (2014). Contributions of pitch contour, tonality, rhythm, and meter to melodic similarity. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2319–2337.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.r-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12057

Schmidt, K. M., & Embretson, S. E. (2003). Item Response Theory and measuring abilities. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of Psychology: Research Methods in Psychology* (pp. 429–446). Hoboken, NJ: John Wiley & Sons.

Schmuckler, M. A. (2009). Components of melodic processing. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology*. New York, NY: Oxford University Press.

Schulze, K., Dowling, W. J., & Tillmann, B. (2012). Working memory for tonal and atonal sequences during a forward and backward recognition task. *Music Perception*, *29*(3), 255–267.

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

Steinbeck, W. (1982). Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse. In *Kieler Schriften zur Musikwissenschaft XXV*. Kassel, Germany: Bärenreiter.

Takeuchi, A. H., & Hulse, S. H. (1992). Key-distance effects in melody recognition reexamined. *Music Perception*, *10*(1), 1–23.

Typke, R., Wiering, F., & Veltkamp, R. C. (2007). Transportation distances and human perception of melodic similarity. *Musicae Scientiae*, *Disc.4A*, 153–181. doi:10.1177/102986490701100107

Uitdenbogerd, A. L. (2002). *Music information retrieval technology*. Doctoral dissertation, RMIT University, Melbourne, Victoria.

Ullén, F., Mosing, M. A., Holm, L., Eriksson, H., & Madison, G. (2014). Psychometric properties and heritability of a new online test for musicality, the Swedish Musical Discrimination Test. *Personality and Individual Differences*, *63*, 87–93. doi:10.1016/j.paid.2014.01.057

Unyk, A. M., Trehub, S. E., Trainor, L. J., & Schellenberg, E. G. (1992). Lullabies and simplicity: A cross-cultural perspective. *Psychology of Music*, *20*(1), 15–28. doi:10.1177/0305735692201002

Vispoel, W. P. (1993). The development and evaluation of a computerized adaptive test of tonal memory. *Journal of Research in Music Education*, *41*(2), 111. doi:10.2307/3345403

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. doi:10.1016/j.lindif.2010.02.004

Watkins, A. J. (1985). Scale, key, and contour in the discrimination of tuned and mistuned approximations to melody. *Perception & Psychophysics*, *37*(4), 275–285. doi:10.3758/BF03211349

Wechsler, D. (2011). Wechsler Abbreviated Scale of Intelligence II. San Antonio, Tex: Psychological Corporation.

Wing, H. D. (1961). *Standardised tests of musical intelligence*. The Mere, England: National

Foundation for Educational Research.

Zenatti, A. (1975). Melodic memory tests: A comparison of normal children and mental defectives. *Journal of Research in Music Education, 23*(1), 41–52. doi:10.2307/3345202

Tables and figures

Table 1

Model fit statistics for the 'same' items

Model	New predictor	*df*	AIC	BIC	log(likelihood)	$\chi^2(1)$	*p*
0	NA	10	4388.6	4451.3	−2184.3	NA	NA
1	Musical training	11	4378.6	4447.6	−2178.3	11.99	< .001
2	Complexity	12	4372.9	4448.1	−2174.4	7.69	.006
3	Tonalness	13	4372.8	4454.3	−2173.4	2.09	.148

Table 2

Model fit statistics for the 'different' items

Model	New predictor	*df*	AIC	BIC	log(likelihood)	$\chi^2(1)$	*p*
0	NA	10	4738.0	4801.2	−2359.0	NA	NA
1	Musical training	11	4679.9	4749.4	−2329.0	60.04	< .001
2	Complexity	12	4677.8	4753.6	−2326.9	4.45	.042
3	Similarity	13	4674.1	4756.3	−2324.1	5.67	.017

Figure 1. Item-wise measures of complexity (composite measure), split by test. The width of each violin plot is proportional to the density of the complexity distribution.

Figure 2. Item-wise measures of similarity (hybrid measure), split by test. The width of each violin plot is proportional to the density of the similarity distribution.

Figure 3. Item-wise measures of tonalness, split by test. The width of each violin plot is proportional to the density of the tonalness distribution.