Learning and recalling melodies: a computational investigation using the melodic recall paradigm.

Sebastian Silas1,2 & Daniel Müllensiefen1,2

1 Goldsmiths, University of London, London, UK

2 Hochschule für Musik, Theater und Medien, Hannover, Germany

Author note

Sebastian Silas has been supported by a doctoral scholarship from the Studienstiftung des deutschen Volkes. This project has been partly supported by funding from the Deutsche Forschungsgemeinschaft (DFG, MU 2722/1-1) awarded to Daniel Müllensiefen. No potential competing interest was reported by the authors. The authors would like to thank Ani Patel and Niels Verosky for their invaluable feedback on this manuscript.

The authors made the following contributions. Sebastian Silas: Writing - Original Draft Preparation, Writing - Review & Editing, Visualization, Formal Analysis, Conceptualization; Daniel Müllensiefen: Conceptualization, Data Collection, Data Pre-Processing, Writing - Original Draft Preparation, Writing - Review & Editing, Formal Analysis.

Correspondence concerning this article should be addressed to Daniel Müllensiefen, Department of Psychology, Goldsmiths, University of London, 8 Lewisham Way, London SE14 6NW, United Kingdom. E-mail: [D.Mullensiefen@gold.ac.uk](mailto:D.Mullensiefen@gold.ac.uk)

Abstract

Using melodic recall paradigm data, we describe an algorithmic approach to assessing melodic learning across multiple attempts. In a first simulation experiment, we show how similarity measures are preferable to assess melodic recall compared to previously-utilised accuracy measures. In Experiment 2, with up to six attempts per melody, 31 participants sang back 28 melodies (length 15-48) presented either as a piano sound or a vocal audio excerpt from real pop songs. Our analysis aimed to predict the similarity between the target melody and participants’ sung recalls across successive attempts. Similarity was measured with different algorithmic measures reflecting various structural aspects of melodies (e.g., tonality, intervallic) and overall similarity. However, previous melodic recall research mentioned, but did not model, that the number of recalled notes increases across attempts, aside from overall performance accuracy, potentially driving overall performance. Consequently, we modelled how the number of recalled notes change alongside similarity. In a mediation analysis, we show how a melody’s length (but not other melodic features) is the main driver of similarity via the number of recalled notes. Therefore, musical features may be secondary to sheer length constraints when learning melodies long enough to require several attempts to recall in full.

*Keywords:* melodic recall, melodic memory, melody learning, working memory, melodic similarity

Learning and recalling melodies: a computational investigation using the melodic recall paradigm.

Most people would not be surprised to hear their friend sing a familiar melody to them, even if their friend were not a great singer or a professional musician. Indeed, remembering melodies is not just an explicitly taught skill useful to professional musicians (Lehmann, Sloboda, & Woody, 2007), but an implicitly acquired ability which most of the general population engage in effortlessly (Bigand & Poulin-Charronnat, 2006; Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Ettlinger, Margulis, & Wong, 2011; Müllensiefen & Halpern, 2014; Schellenberg, Weiss, Peng, & Alam, 2019; Tillmann, Bharucha, & Bigand, 2000). Consequently, for most people in the general population, melodic memory encoding and retrieval processes are a normal part of life, even though for many such abilities are only implicitly acquired and exercised, rather than formally trained (Lehmann et al., 2007). In some basic respect, remembering and recalling melodies could be viewed a general skill, not dependent on formal music training or expertise.

The purpose of this study is to model such melodic memory and recall processes in a quantitative way, and to understand how mental representations of melodies develop over short periods of time, after repeated exposure to the same melodic target stimulus. Specifically, we advocate the *melodic recall paradigm* (Sloboda & Parker, 1985), and in doing so, like Okada and Slevc (2021) recently argued, among others (Buren, Müllensiefen, Roeske, & Degé, 2021; Hallam & Creech, 2010), emphasise the importance of musical *production* tasks to gaining a comprehensive understanding of musical abilities. However, we take modelling of the melodic recall paradigm forward in two main respects. First, we reason for and employ similarity metrics to score melodic recall data, and argue that simple accuracy measures alone are not sufficient to understand melodic recall. Second, whilst Sloboda and Parker (1985) noted that participants gradually attempt to sing more notes across each consecutive attempt at recalling the same melody, they did not formally model such changes across attempts. We contend that not formally modelling the change in number of recalled notes per attempt is a fundamental omission in previous melodic recall studies (e.g., Ogawa, Kimura, & Mito, 1995; Sloboda & Parker, 1985; Zielinska & Miklaszewski, 1992). In particular, we suggest that modelling the change in number of recalled notes per attempt, in parallel to the overall change in overall performance, offers at least three main advantages to melodic recall research. First, it points to the potential application of models and ideas from already well-established theories of produced mental representations in non-musical domains [e.g., serial recall; Anderson (1972)]. Second, it reminds us that there are general constraints on human memory (Christiansen & Chater, 2016; Cowan, 2010; Miller, 1956; Oberauer & Cowan, 2007), and that not all variance in melodic recall behaviour may be explained in musicological terms, which may suggest a weighting of explanatory focus from domain-specific to domain-general memory mechanisms. Lastly, it enables key insights into how the encoding of melodic information works, which may otherwise be lost in statistical inference that does not take into account domain-general memory faculties (see Silas, Müllensiefen, Gelding, Frieler, & Harrison, 2022 for a related discussion).

In summary, as this paper will show, concepts related to general working memory constraints (e.g., non-musical serial recall, item length) appear to be more important to explaining melodic recall than any other musicological considerations (e.g., interval representations, tonality) when the length of the melodies certainly requires multiple attempts to sing back all the notes (e.g., length 15-48), or at least with the data presented here. In other words: “melodic recall” appears to be closely related to “recall” in other memory domains.

## Music and Working Memory

The construct of working memory is now well-developed in psychology, with the most popular model being Baddeley and Hitch (1974)’s multi-component model, subsequently updated in Baddeley (2000). Working memory refers to the ability to transform and manipulate information in short-term memory. In general, it is thought to comprise components for manipulating phonic and visual stimuli separately. Music scholars have long recognised the important role of working memory in musical behaviours, particularly those involving aural skills (Chenette, 2021; Cornelius & Brown, 2020; Gates, 2021; Karpinski, 2000). Indeed, those with formal music training have widely been documented to have better working memory capacities (Talamini, Altoè, Carretti, & Grassi, 2017; Talamini, Carretti, & Grassi, 2016).

It has been argued by some (e.g., Berz, 1995) that general working memory models do not explain working memory for musical stimuli well. Other authors such as Ericsson and Kintsch (1995) contend that the development of expertise, in specialised domains such as formal music training and chess, cultivates domain-specific forms of working memory, which they refer to as “long-term working memory”, whereby (musical) abilities are subserved by relatively specialised systems, quite distinct from general working memory. In our own previous research, we have documented the possible scenarios which might explain the links between domain-general and domain-specific (music) working memory faculties: they may be relatively (statistically) disparate, but nonetheless, rely on each other, potentially bidirectionally (Silas et al., 2022). The implications of this are that, perhaps by definition, musical abilities are subserved by both domain-general (potentially to do more with inherited characteristics) and domain-specific (potentially more to do with training) faculties. In other words, someone with a very good general working memory might be able to demonstrate a similar level of (e.g., sung recall) performance to someone who has had more musical training. The former’s general faculties may help them monitor their performance as well as someone who has carved out music-specific templates to aid the same task. The underlying processes may be different, but the observable phenotype similar. Framed in terms of our study: if music conforms to a style that people in the general population are familiar with, do musical features (often better remembered by expert musicians) tend to matter? With relatively simplistic, familiar musical styles, is performance really mediated by music-specific processes, or could it be more domain-general processes which turn out to be important? If melodies are long enough to require multiple attempts to sing in full, are musical features beyond length particularly important, compared to the length of a melody alone? Next, we discuss previous approaches to studying melodic memory.

## Melodic Recognition Paradigm

Traditionally, melodic memory has been investigated frequently using different variants of the *melodic recognition paradigm* (Idson & Massaro, 1978). In this paradigm, the listener hears a melody in a training phase and then a second melody in a test phase. The second can be identical, similar in some musical sense, or completely different from the first (e.g., Dowling & Fujitani, 1971). The participants’ task is to tell whether the two melodies are identical or not. The rationale of this paradigm is that undetected differences between two melodies reflect differences in musical dimensions which are not retained in memory, or are forgotten easily. Differences that participants do detect are supposed to happen in a musical dimension which is represented in memory (for a good and compact description of the paradigm see e.g. Idson and Massaro (1978), p. 554). In many such studies, melodies used as stimuli were composed and/or manipulated by the experimenters to show the desired differences in the specific musical dimensions. Such studies show, for example, that, at least under certain conditions, contour representations of melodies are more easily retained in memory than interval representations (Dowling, 1978; Dowling, Kwak, & Andrews, 1995; Edworthy, 1985; Massaro, Kallman, & Kelly, 1980), shorter sequences are recognized better than longer ones (Edworthy, 1985; Long, 1977), and after short retention intervals, contour is retained better, but after long retention interval memory performance for tonality and intervallic information is superior (Dewitt & Crowder, 1986; Dowling, 1991; Dowling & Bartlett, 1981).

There are two main disadvantages of the melodic recognition paradigm for the study of melodic memory: Firstly, participant responses are limited to a binary decision (i.e. ‘identical’ vs ‘not identical’), possibly with a confidence judgement on an ordered scale. This has been criticized for discarding a lot of information that may be relevant in analyzing the actual memory representations, which are presumably much richer than such a binary decision can reflect. Secondly, the experimental melodies and their according variants are, in most cases, artificially constructed to fulfill the constraints of the experimental design. This often results in the usage of pitch sequences that are stylistically unfamiliar to participants and may be rarely encountered in actual human melodic processing. If realistic musical material is used, differences between the to-be-compared excerpts introduced by the experimenter can often appear obvious or artificial. Subtle differences and naturally occurring nuances between memory representation and the original may thus remain undiscovered (e.g., Kauffman & Carlsen, 1989).

Recent developments to the related experimental approach of melodic discrimination testing via explanatory item response theory (Harrison, Collins, & Müllensiefen, 2017; Harrison, Musil, & Müllensiefen, 2016) and usage of large-scale musical corpora (e.g., Baker, 2021; Pfleiderer, Frieler, Abeßer, Zaddach, & Burkhart, 2017) have bolstered and improved some of these aspects of the melodic recognition paradigm. However, the so-called melodic *recall* paradigm, employed in this study, offers a different kind of insight into the different musical dimensions retained in memory.

## Melodic Recall Paradigm

There have been a few studies employing the melodic recall paradigm to investigate memory for melodies, with Sloboda and Parker (1985) probably being the most well-known. Sloboda and Parker (1985) played the 30-note instrumental melody of a folk song to participants and asked them to sing back whatever they remembered from the melody. As participants found it very difficult to sing back much of this relatively simple and comparatively short folk tune, they could play the melody up to six times, with a chance to sing back the melody again after each hearing. As a result, a sung recall for every trial attempt and every participant was obtained. With a manual but quasi-algorithmic analysis technique, Sloboda and Parker (1985) showed that the (phrase, metric, harmonic) structure of the heard melody was learned rather early in attempts while intervallic and rhythmic details stayed quite inaccurate until later attempts. We note the operalisation and assumption of this approach, which we follow here: improvements in sung recall are taken as evidence of learning a melody. In other words, to improve on singing back a melody, it is necessary that the melody has been remembered (i.e., learned) better on each consecutive attempt. We use the term “learning” throughout the manuscript, specifically referring to the task at hand, which is sung recall, as distinct from other tasks (e.g., aural dictation) but do not intend to suggest that the melody is necessarily learned beyond the task of singing.

Sloboda and Parker (1985) observed that the sung recalls got considerably longer over the six repetitions, but the ratio between the number of correctly recalled notes and the overall number of sung notes stayed approximately constant. Among the 48 trials of the eight participants they tested, they observed not a single rendition without error. In their discussion, they concluded that, in accordance with the notion of generative grammars for melodies, melodic structure seems to be a feature that is preferably abstracted in memory, while details such as exact pitches and durations are rather improvised within the constraints of the melodic structure retained in memory. Sloboda and Parker (1985)’s results were partially reconfirmed by Müllensiefen and Wiggins (2011) who used Sloboda and Parker (1985)’s original transcribed recall data, but employed a computational approach to analysing it. Their algorithmic approach suggested some different interpretations of the data. For instance, Sloboda and Parker (1985) observed no increase in performance across attempts, which is surprising, because it suggests that melodic features are not incrementally extracted through repeated exposure. However, Müllensiefen and Wiggins (2011) presented evidence of learning: participants seem to be able to recall the melody better across repeated attempts, as indicated by increases in *similarity* (not accuracy) between the sung recall and the target melody. This suggests that accuracy alone may not be an appropriate measure of melodic recall performance.

A few studies after Sloboda and Parker (1985) followed the same experimental approach of using a melodic recall paradigm, but differed in their use of experimental materials [more melodies, tonal vs. modal melodies; Oura and Hatano (1988)], participants [more participants, participants with and without absolute pitch or formal musical background; Ogawa et al. (1995); Zielinska and Miklaszewski (1992)], and number of trials per participant and melody (up to 10). The error rates over trials that Zielinska and Miklaszewski (1992) obtained suggest that participants can reach a level of almost error-free recalls if they are given enough trials, and that there is a particular point where the relative errors in the sung recalls start to diminish more noticeably. The position of this point seems to depend primarily on the amount of musical training of the participants. With music students possessing absolute pitch, the point is already at the second trial, whereas music students without absolute pitch need four repetitions before overall error rates decrease. The fact that there is a particular point where participants’ error rates significantly start decreasing does not necessarily speak against Sloboda’s and Parker’s claim that melodic structure is acquired first. Zielinska and Miklaszewski (1992), as well as Oura and Hatano (1988), also discovered that their participants first memorised the structure by segmenting the melodic stream into ordered phrases and improvising on details. This was especially true for the formally trained participants, while music novices tended to commit rather ‘unmusical’ errors, such as modulations to different tonalities or errors on phrase contours, as Oura and Hatano (1988) note. For a more thorough review of general memory paradigms and their adaptation for melodic memory research, we refer the reader to Müllensiefen and Wiggins (2011).

# Methodological Issues with Melodic Recall Research

Despite the possibility of giving interesting insights into the mechanisms of memory for melodies, the melodic recall paradigm as applied by the cited studies has some inherent problems. Firstly, previous cited studies using the melodic recall paradigm relied on a hand-made comparison analysis between target melody and sung recalls. Consequently, the number of recalls to be analysed was limited. For example, Sloboda and Parker (1985) analysed 48 renditions from their participants, while Oura and Hatano had 320, and Zielnska and Miklaszewski (1992) 310 renditions to base their analyses on. For going beyond this level of analysis, the computer suggests itself as an aid (i.e., algorithmic analysis). The computer-based analysis used in the present study allows us to cope with around 2,250 sung recalls. In turn, this higher number of melodic objects allows the deployment of techniques from statistical modeling, which require many data points to be used effectively. Secondly, previous methods to assess the quality of a sung recall involved accuracy-based measures, which alone are inadequate for meaningfully assessing melodic recall behaviour, as we will demonstrate. Third, whilst Sloboda and Parker (1985) noted that sung recalls got longer (though not necessarily better) over subsequent trials, they did not model this effect. This methodological omission leads to a theoretical one: it neglects to observe the domain-general aspects of sung recall, which do not differ from normal recall.

## The Present Study: Methodological Advances for Melodic Recall Research

To meet these shortcomings of previous melodic recall research, we make a number of methodological advances. These are to: i) employ an algorithmic analysis of melodic recall data; ii) employ similarity metrics over (but including) accuracy measures to score such data; and iii) to model the change in number of recalled notes across attempts in addition to changes in melodic similarity, which allows comparisons to general memory faculties.

To that end, we conduct two experiments. Experiment 1 is a simulation study where we describe and compare accuracy and similarity-based measures. Formally, we show how accuracy and similarity measures diverge for different simulated conditions. This highlights the different properties of similarity-based measures and suggests their superiority over accuracy measures for melodic recall data. In Experiment 2, we present an experiment using melodic recall data collected from human participants. Here, we focus on another important point that has been overlooked in melodic recall research: the lack of a statistical model to support how the number of recalled notes changes across consecutive attempts and how this changes alongside overall performance, as measured by melodic similarity.

# Experiment 1: A comparison of accuracy vs. similarity measures applied to melodic recall data

The aim is of Experiment 1 is to succinctly explore the difference between accuracy vs. similarity measures. We explore how they diverge in a quantitative manner when applied to assessing how well a sung recall matches a target melody. We first operationalise them both verbally and mathematically, also giving some reference to real musical examples. Subsequently, we proceed with a simulation study using the stimuli set that we use with real participants in Experiment 2, to evaluate how their properties change as a function of various experimental manipulations.

# Operationalising Accuracy vs. Similarity

## What is accuracy?

Traditionally, melodic recall data has been scored with accuracy measures (Sloboda & Parker, 1985). In the context of melodic recall, broadly speaking, accuracy can be defined as the number of correct notes relative to something else. That something else could be i) the number of unique notes in the target stimulus or ii) the number of recalled notes. Let us call a note that is both contained in the target melody and the sung recall a “hit”, a note recalled that is not in the target stimulus, a “false alarm” and a note in the target melody but not the sung recall, a “miss”.

*Accuracy* can be defined as:

Another measure related to accuracy is *precision*, defined as:

which means the denominator will be the number of notes in the stimulus and is, hence, the proportion of notes in the target stimulus recalled. This is the measure that Sloboda and Parker (1985) utilised on sung recall data.

Lastly, *recall* is defined as:

where the denominator represents the number of notes in the sung recall.

Precision and recall can be combined together into a score known as the F1 score, which is the harmonic mean of precision and recall, simplifying to:

The F1 score as well as precision and recall have been used extensively in computational contexts and music information retrieval contexts (Pearce, Müllensiefen, & Wiggins, 2010). One issue in melodic recall research is that the length of a recall can be substantially different from that of a target melody length, in terms of number of notes. One way of dealing with this, in the context of accuracy, is to normalise based on the length of the stimuli:

where Accuracy is computed via Eq. 1 and LogNormal represents a log normal transformation function to map the accuracy value into the closed unit interval range.

We profile these measures of accuracy, *Accuracy*, *Recall*, *Precision*, *AccuracyAdjusted* and *F1*, in Experiment 1.

## Methodological Advance for Melodic Recall Research

However, as highlighted by Sturm (2013)’s paper title, “classification accuracy is not enough” when it comes to assessing musical information. As argued there, this is because accuracy measures do not work at a high enough level to meaningfully represent musical structures. In our case - melodic recall - note-by-note accuracy measures fail to capture important aspects of musical structure, such as interval patterns and the specific order of notes. In this way, accuracy is not particularly meaningful in a musical context, and we suggest is an inadequate measure of sung recall.

This can be viewed as a limitation of previous melodic recall research. The approach taken by Sloboda and Parker (1985) and others were construed from reasonable musicological considerations, but do not represent the only way to compare two melodic objects. Moreover, such comparisons have not been tested for their ecological validity. As noted, Sloboda and Parker (1985) used the *precision* measure of accuracy. However, measuring note-by-note reproduction accuracy hardly reflects the fact that the original and recall may differ in many improvised notes, but that on other levels of human melodic understanding the sung rendition might be “highly related to the original in many respects” (Sloboda & Parker, 1985, p. 159). Thus, Sloboda and Parker recognised in 1985: “there is no theory of melodic identity”.

### Melodic Similarity.

To overcome this issue, we introduce the notion of a melodic similarity metric as being more appropriate for scoring melodic recall data. In essence, measures of melodic similarity embody important notions of melodic identity into them. The concept of similarity is hence more sophisticated than simple note-by-note accuracy measures and capture more musical variance. In the scientific area that has been termed *Music Information Retrieval* (Downie, 2003), and that has seen a large boost in recent years, several approaches to *similarity* measurement for melodies and other musical objects have been developed (e.g., Müllensiefen & Frieler, 2004a; Pearce & Müllensiefen, 2017; Typke, Wiering, & Veltkamp, 2007; Yuan et al., 2020). The similarity measures employed here are favoured because they proved their effectiveness and ecological validity (or rather comparability) with the notion of melodic similarity of musically experienced participants in separate studies (Müllensiefen & Frieler, 2004b, 2007; Müllensiefen & Pendzich, 2009). Therefore, while there still might not be an undisputed theory of melodic identity, as Sloboda and Parker claimed in 1985, this study will use at least some algorithms that came quite close in emulating musically experienced participants’ similarity judgements.

Having obtained numerical representations of both the sung recalls and the target melodies, the similarities between a target melody and sung recall of that melody, for each attempt of each participant, can be calculated using the algorithmic similarity measures described in Müllensiefen and Frieler (2004b). The similarity measures we employed complied with two main points already raised by Sloboda and Parker (1985) in the discussion of their methodology of melodic comparison: in most cases - that is especially true for the earlier trials - participants only recall a smaller part of the original melody, which may not even start with the beginning of the original. Thus, a similarity measure (or algorithm) must be chosen that automatically looks for the best alignment of the (short) melodic sequence of the sung recall with the original melodic sequence. An algorithm for the optimal alignment of two symbol sequences that has been widely used in domains such as text retrieval or bio-computing, as well as music information retrieval, is the so-called Edit Distance or Levenshtein distance (e.g., Mongeau & Sankoff, 1990). The Edit Distance is the minimum number of operations it takes to transform one symbol string into another: the possible operations being insertion, deletion, and substitution. The actual calculation of the Edit Distance is carried out using dynamic programming and is not explained here. For a general reference regarding the algorithm see e.g., Gusfield (1997). In this case, the maximal Edit Distance of two strings is equal to the length of the longer string. To convert the Edit Distance into a similarity measure with a range of values we use the following:

where and denote the element counts of strings and respectively, and stands for the Edit Distance between strings and .

Just like the manual scoring techniques employed by Sloboda and Parker (1985), the edit distance calculates the similarity between two symbolic sequences by taking the number of edits (i.e. additions, deletions or substitutions) that are necessary to transform one of the sequences into the other, and dividing this figure by the number of symbols in the longer sequence. It thus could be argued that Sloboda and Parker intuitively used a version of the edit distance, evaluating the similarity between the recalls of their participants on the original melody, keeping the order of notes in mind.

However, instead of applying edit distance to raw note values, here the edit distance is computed on various symbolic representations of musical dimensions [i.e., relative pitch sequences - intervals - as opposed to single pitches; rhythm sequences; and sequences of implied harmonies; Müllensiefen and Frieler (2004b)]. Specifically, we employ the *opti3* measure of melodic similarity (Müllensiefen & Frieler, 2004b) as our main dependent variable. *opti3* is a hybrid measure derived from the weighted sum of three individual measures which represent different aspects of melodic similarity. The similarity in interval content is captured by the *ngrukkon* measure is based on the Ukkonen Measure that measures the difference of the occurrence frequencies of interval trigrams () contained within the target melody () and the comparison melody () (see Uitdenbogerd (2002). Formally:

As the Ukkonen Measure is a distance measure in its original definition, we normalise by the maximum possible number of -grams and subtract the result from 1:

Note that the Ukkonen measure is not based on the edit distance but still takes order of notes into account at a local level by comparing trigrams.

Harmonic similarity is measured by the *harmcore* measure. This measure is based on the chords implied by a melodic sequence, taking pitches and durations into account. Implied harmonies are computed using the Krumhansl-Schmuckler algorithm (Krumhansl, 1990) and the harmonic sequences of the two melodies are compared by computing the number of operations necessary to transform one harmonic sequence into the other sequence via the edit distance. Finally, likewise, rhythmic similarity is computed by first categorizing the durations of the notes of both melodies (known as “fuzzification”) and then applying the edit distance to measure the distance between the two sequences of categorised durations. The resulting measure of rhythmic similarity is called *rhythfuzz* (Müllensiefen & Frieler, 2004b). Note that *rhythfuzz* does not take metric information into account and works solely on the basis of (relative) note durations. Similarly, *ngrukkon* works with interval information and is hence invariant to transposition.

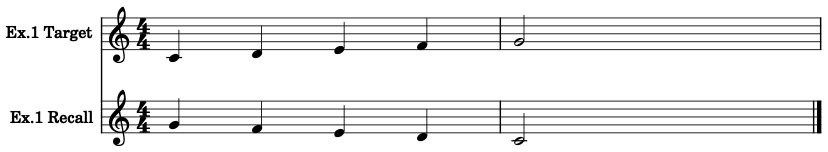
Based on the perceptual data collected by Müllensiefen and Frieler (2004b), the three individual measures are weighted and combined to form a single aggregate measure of melodic similarity, *opti3*. Hence, *opti3* is sensitive to similarities and differences in three important aspects of melodic perception (pitch intervals, harmony, rhythm). We note that all three individual measures (*ngrukkon*, *harmcore*, *rhythfuzz*) can take values between 0 (= no similarity) and 1 (= identity) and are length-normalized by considering the number of elements of the longer melody. *opti3* then comprises (Müllensiefen & Frieler, 2004b):

Note that we here present the normalised weights, which constrain the values to be [0,1]. We note that *opti3* is dependent on the length of the two comparison melodies in only a soft sense, which is particularly relevant to Experiment 2 of this paper, where we use the number of recalled notes as an auxiliary dependent variable. If one melody is shorter than the other, at least some of the melodic identity is destroyed: necessarily, the rhythmic (*rhythfuzz*) and intervallic (*ngrukkon*) components, but not necessarily the harmonic (*harmcore*) component. It should hopefully be clear that *opti3* captures far more (musical) information than melody length(s) alone and/or accuracy-style measures. The ecological validity of the aggregate similarity measure has been established in several perceptual experiments (Müllensiefen & Pendzich, 2009; Yuan et al., 2020). See Appendix A to compare the similarity measures and Appendix B for notated examples of development in sung recall performance and a qualitative description of their change in similarity.

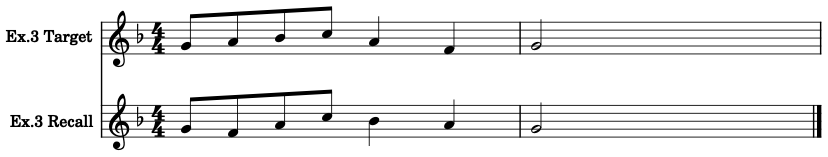
### What is the difference between accuracy and similarity measurements?.

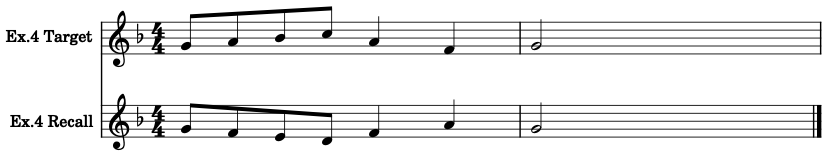
As should hopefully be clear, similarity measures pay more attention to musical features that arise from the relationships between pitch and rhythmic values. They could be considered more global in nature. Conversely, note-by-note accuracy measures only pay attention to the very local level of notes, which does not respect the emergent properties of musical features. Consequently, similarity has a more perceptual quality. In this way, similarity algorithms have been used to predict subjective similarity judgements, for example, in musical plagiarism court cases, with excellent success (Müllensiefen & Pendzich, 2009; Yuan et al., 2020).

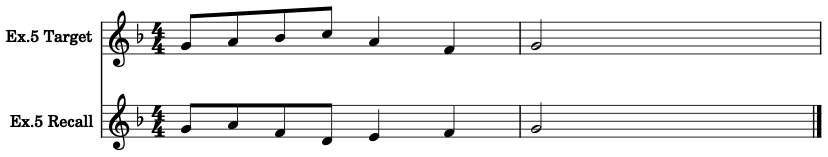
We now score some simple examples to demonstrate the difference between accuracy and similarity (see Table 1 and Figure 1). Let us take the first five notes of the C major scale as our target. If the target and recall are the same, all measures of accuracy and similarity will be 1. However, as Example 1 shows, if the pitch order is reversed, the accuracy measures will be 1, but the overall measure of similarity (*opti3*), nearly half this. This corresponds to the fact that the notes are the same, but the order highly distorts the melodic identity (note that the harmonic and rhythmic identity is preserved, however). Conversely, as Example #2 shows, transposing the recall by a semitone causes a large deterioration in the accuracy scores, since nearly all notes are different, but the similarity measure is 1, corresponding to the fact that the melodic identity is preserved under the principle of transposition invariance. For Examples 3-6, we take the humourous example of the “The Lick” - the archetypal jazz pattern made famous by YouTube[[1]](#footnote-29) and transform it by the familiar melodic transformations of retrograde, inversion and retrograde inversion. Example #3 shows that the retrograde preserves accuracy but destroys identity and similarity. Examples #4 and #5 show that the inversion and retrograde inversion somewhat preserves accuracy but destroys identity and similarity. Finally, consider no pitch value changes, but only rhythmic changes, which destroys rhythmic identity and lowers overall melodic similarity (*opti3*), but preserves accuracy.











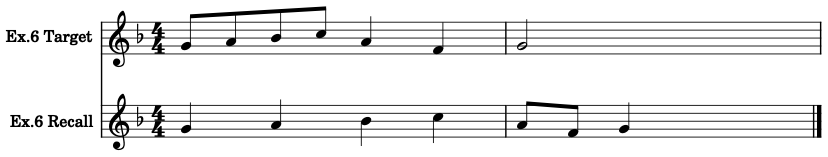


Figure1: Notation of sung recall comparison examples

Table 1: Examples of two comparison melodies scored for their level of similarity.

| Example No. | Example Desc. | Acc. | F1 | AccAdj | opti3 | ngrukkon | rhythfuzz | harmcore |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Reversed Recall | 1.00 | 1.00 | 1.00 | 0.51 | 0.00 | 1.00 | 1 |
| 2 | Transposed Recall | 0.11 | 0.20 | 0.20 | 1.00 | 1.00 | 1.00 | 1 |
| 3 | The Lick vs. its retrograde (rhythms the same) | 1.00 | 1.00 | 1.00 | 0.51 | 0.00 | 1.00 | 1 |
| 4 | The Lick plus its inversion (rhythms the same) | 0.56 | 0.71 | 0.71 | 0.27 | 0.00 | 1.00 | 0 |
| 5 | The Lick plus its retrograde inversion (rhythms the same) | 0.56 | 0.71 | 0.71 | 0.40 | 0.25 | 1.00 | 0 |
| 6 | The Lick plus its rhythm values changed. | 1.00 | 1.00 | 1.00 | 0.66 | 1.00 | 0.14 | 1 |

The examples should intuitively demonstrate how accuracy and similarity measures diverge and specifically, that similarity metrics embody notions of perceptual similarity across relevant musical dimensions, as opposed to note-by-note accuracy measures, which do not necessarily embody anything musical into their consideration. In other words: accuracy measures may score a target and its recall very highly, despite the musical differences being very large, and the opposite, a very musically similar target and recall very badly! These observations are not new. Perceptual experiments have demonstrated that the familiar transformations of retrograde, inversion and retrograde inversion destroy melodic identity (e.g., Dowling, 1972). However, we are not aware of such a distinction being made clearly within melodic recall research. To profile the difference more quantitatively, we now run a series of short simulation studies.

# Method

The method taken here is to start with the stimuli set of melodies that we use with real participants in Experiment 2. Starting with this set, we transform the melodies in various ways so that they become more or less like their originals. We then compare how melodies vary on accuracy and similarity scores as a function of the number of transformations.

## Materials

14 pop songs were used as the basis for the experimental material. Two different melodies were taken from each song to form a stimuli set of 28 melodies in total (9–21 seconds; length = 15-48 notes; see Appendix C). Melodies were selected to represent a wide range of Western popular music styles, ranging from easy listening ballads and Schlager via mainstream pop, rock, and disco to blues, R’n’B and hip hop. They were taken from hit song collections covering repertoire form the 1960s to 2000, but overly popular or well-known songs were avoided. The set of melodies was selected to provide a large range of stylistic features within the context of Western popular music. This includes different melodic and singing styles as well rhythms and meters that are idiomatic for certain genre. This wide stylistic breadth should allow for a better generalisation across Western popular music than a more homogeneous stimulus set that exclude certain musical features deliberately.

## Simulation Experiments

We conducted a series of eight simulation experiment using the stimuli described above. In each experiment, we manipulated each stimulus via transformations either on the raw pitch classes or the duration values. Subsequently, all results were aggregated across melodies. We describe each sub-experiment succinctly in Table 2 below. For more detailed information, please consult our experiment code and data (i.e., simulated melodies)[[2]](#footnote-55).

Table 2: Description of experiments simulating human errors on a sung recall task

| No. | Name | Description |
| --- | --- | --- |
| 1A | Rhythmic jitter | Duration values were jittered by various amounts (0, 0.01, 0.10, 1, 2, 5), using the jitter *R* function. This corresponds to singing the rhythms but not pitches incorrectly. |
| 1B | Pitch insertions | Various number of notes were randomly inserted, corresponding to cases where human participants mistakenly add random notes to their sung recalls. |
| 1C | Pitch deletions | Notes were randomly deleted, corresponding to cases where human participants mistakenly miss notes in their sung recalls. |
| 1D | Pitch substitutions | Notes were randomly substituted (at any location in the melody), corresponding to cases where human participants sing some random notes wrong in their sung recalls. |
| 1E | Combined pitch insertions, deletions, and substitutions | The last three experiments combined, corresponding to human participants making various random mistakes. |
| 1F | Combined pitch insertions, deletions, and substitutions and rhythmic jitter | Rhythmic jitter and pitch transformations as described above (i.e., experiments 1A and 1E) were transformed simultaneously, corresponding to singing pitches and rhythms wrong. |
| 1G | Length mismatch | Notes were removed from the end of the target melody to create a length mismatch between target and recall, corresponding to a participant not yet being able to sing an entire melody back. |
| 1H | Scramble | Different sized chunks of the melody were scrambled, such that the same notes were in each chunk, but the order of notes was changed randomly. This corresponds to human participants retaining a gist of the melodic identity, but not a precise representation of its structure. |

# Data Analysis

First, we assess the descriptive statistics for the variables and, in particular, we compared the coefficient of variation (*CV*) between our measures. The *CV* is not the same as the well-known coefficient of determination (R2). Instead, the *CV* is the standard deviation (SD) of a measure divided by its mean. The *CV* is preferred over the standard deviation for comparing variance across measures, because the *CV* is a dimensionless number, and hence, facilitates comparison of the SD across different datasets or measures, with different means. By comparing the *CV* across measures, we can assess whether some capture more variance than others.

Next, we inspect each sub-experiment results via graphs, where the y-axis always represents the score on each accuracy or similarity measure and the x-axis represents a function of transformations. The specific transformations are operationalised in Table 2.

Finally, we formally model all experimental data simultaneously. The dependent variable is the (accuracy or similarity) score; Measure (*accuracy*, *opti3* etc.) and Experiment (1A-IH) are categorical predictors; length of sung recall is a numeric predictor as well as the interactions between i) Measure and length of sung recall and ii) Measure and Experiment were included as predictors. There were 35,371 observations. To facilitate interpretation of the model parameters, sung recall length was standardised before model fitting (note the other parameters are already [0, 1]).

# Results

Table 3 presents descriptive statistics. The similarity measures have the highest coefficient of variation compared to the accuracy measures, suggesting that they capture more variance than accuracy measures, at least in the context of our experiment. This is also suggested by the graphs in Figure 2, whereby similarity measures tend to degrade more than accuracy measures, as a function of the musical errors we simulated. The other results shown in Figure 2 can be summarised as follows: 1A) adding jitter to duration values causes *rhythfuzz* to degrade as well as *harmcore* (because it offsets the alignment of harmonic sequences), and consequently, *opti3*. All accuracy measures are unaffected. 1B) note insertions cause all measures to degrade, except accuracy and precision, which stay constant. Similarity measures generally degrade more prominently than accuracy-style measures, particularly *ngrukkon*; 1C) note deletions cause all measures to degrade, except *accuracy* and *recall*, which stay constant. Similarity measures generally degrade more quickly than accuacy-style measures, particularly *ngrukkon*; 1D) note deletions cause all measures degrade, except *rhythfuzz*, whichs stay constant. The other similarity measures (*ngrukkon*, *harmcore*, *opti3*) degrade more prominently than accuracy-style measures, particularly *ngrukkon*. 1E) combined pitch transformations (insertions, deletions and substitutions) cause all measures to degrade, but similarity measures more prominently than accuracy-style measures; 1F) combined pitch transformations (insertions, deletions and substitutions) and rhythmic jitter cause all measures to degrade, but similarity measures more prominently than accuracy-style measures; 1G) as the length of the recall increases towards the length of the target melody, all measures increase, except recall and accuracy, which are always 1; 1H) Scrambling the order of pitches affects no accuracy measures or *rhythfuzz*. All other similarity measures deteriorate as a function of scrambling.

Table 3: Descriptive statistics for all simulation experiment results, ordered by the coefficient of variation.

| Measure | Mean | SD | Coefficient of Variation |
| --- | --- | --- | --- |
| harmcore | 0.50 | 0.40 | 0.80 |
| ngrukkon | 0.37 | 0.27 | 0.74 |
| opti3 | 0.44 | 0.25 | 0.56 |
| rhythfuzz | 0.68 | 0.30 | 0.44 |
| accuracy\_adjusted | 0.73 | 0.25 | 0.34 |
| accuracy | 0.78 | 0.20 | 0.26 |
| recall | 0.82 | 0.18 | 0.22 |
| F1\_score | 0.86 | 0.15 | 0.17 |
| precision | 0.95 | 0.13 | 0.14 |

![Figure 2: Simulation experiment results for accuracy vs. similarity measures.](data:application/pdf;base64,)

Figure2: Simulation experiment results for accuracy vs. similarity measures.

Table 4 shows the general linear regression model results. All similarity measures tend to have lower scores, as represented by all negative beta coefficients (). Accuracy measures have higher beta coefficients. This corresponds to our descriptive results, whereby accuracy measures were less likely to be affected by transformations (simulated human errors) because they do not respect note orders, for example, so tend to score more highly. The experiment/transformation type differentially affects the accuracy similarity scores, as represented by nearly all significant p-values (*p* < .001), suggesting that different measures are more sensitive to some transformations than others. The transformation associated with overall lowest scores was substitutions () and highest scores are rhythmic transformations . Sung recall length is generally associated with a higher score () (i.e., scores tend to be higher if a recall is closer to its target in length). *accuracy\_adjusted*, *ngrukkon*, *precision* and *recall* have a significant interaction with sung recall length, but F1 score, *harmcore*, *opti3* and *rhythfuzz* do not. The model had a R2 value of 0.58 (adjusted = 0.58).

Table 4:

*Regression model regressing score onto measure, sung recall length and experiment, plus interactions (interactions not displayed).*

| Predictor |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 0.72 | [0.70, 0.74] | 56.98 | 35290 | < .001 |
| Measureaccuracy adjusted | -0.29 | [-0.33, -0.26] | -16.45 | 35290 | < .001 |
| MeasureF1 score | 0.06 | [0.02, 0.09] | 3.20 | 35290 | .001 |
| Measureharmcore | -0.23 | [-0.26, -0.19] | -12.07 | 35290 | < .001 |
| Measurengrukkon | -0.40 | [-0.44, -0.37] | -21.46 | 35290 | < .001 |
| Measureopti3 | -0.37 | [-0.41, -0.33] | -19.72 | 35290 | < .001 |
| Measureprecision | -0.02 | [-0.05, 0.02] | -1.05 | 35290 | .294 |
| Measurerecall | 0.29 | [0.26, 0.33] | 15.91 | 35290 | < .001 |
| Measurerhythfuzz | -0.22 | [-0.26, -0.18] | -11.75 | 35290 | < .001 |
| Experimentflip out notes | -0.06 | [-0.09, -0.02] | -3.16 | 35290 | .002 |
| Experimentflip out notes and rhythm jitter | -0.06 | [-0.09, -0.03] | -4.49 | 35290 | < .001 |
| Experimentinsertions | -0.11 | [-0.15, -0.08] | -5.89 | 35290 | < .001 |
| Experimentlength mismatch | 0.09 | [0.06, 0.13] | 5.55 | 35290 | < .001 |
| Experimentrhythm jitter | 0.23 | [0.19, 0.27] | 11.17 | 35290 | < .001 |
| Experimentscramble | 0.22 | [0.19, 0.25] | 14.36 | 35290 | < .001 |
| Experimentsubstitutions | -0.15 | [-0.18, -0.12] | -8.47 | 35290 | < .001 |
| Sung recall length | 0.20 | [0.14, 0.27] | 6.36 | 35290 | < .001 |

# Discussion

Our simulation studies generally show that accuracy and similarity measures produce different assessments of musical behaviour, such as sung recall. This is not surprising, since similarity measures make comparisons on higher-order musical features. As our examples showed intuitively, accuracy measures are not ideal for assessing musical behaviour. This is because they do not have notions of melodic identity embodied in them. Conversely, melodic similarity measures explicitly embody important musical dimensions, and make comparisons on such dimensions. The particular dimensions used by our main dependent variable, *opti3* are based on human similarity judgements (e.g., Müllensiefen & Frieler, 2007), giving them ecological validity - a property that accuracy measures do not have in the context of sung recall. Additionally, as demonstrated, melodic similarity, as derived here, captures more variance as a function of various simulated musical errors, which is an additional useful property. This is because accuracy measures do not capture some forms of musical errors, like singing notes in the wrong order. Conversely, this would affect measures like *ngrukkon*, which assesses interval sequences, but not *harmcore*, which captures overall harmony. In other words, different domains are measured simultaneously by *opti3*. Since someone can be musically more or less accurate in different respects, the measure may be both punitive or benovelent on each dimension, but it respects the fact that musical ability is multidimensional: one may sing a rhythm wrong, but the notes right; the harmony wrong but the rhythms correct. For these reasons, we proceed with Experiment 2 using *opti3* as our main dependent variable to measure overall melodic (sung) recall performance.

# Experiment 2: How do we learn melodies? A melodic similarity-based perspective.

The aim of Experiment 2 is to employ the recall paradigm in an experiment with real participants, much the way that aforementioned studies (e.g., Ogawa et al., 1995; Sloboda & Parker, 1985) have used it. However, this study takes steps ahead in comparison with previous studies using the melodic recall paradigm in at least five basic aspects: (i) The number of different melodies presented: 14, still not large, but substantially larger to previous research which used 1-2 melodies as targets; (ii) The overall number of overall sung recalls to be analysed (around 2,250); (iii) The usage of unambiguously defined and thoroughly tested algorithms of melodic similarity for various musical dimensions; (iv) The modeling of participant responses in statistical models that allows an interpretation of memory mechanisms that involve music-structural variables as well as variables concerning the experimental design and participants’ musical background. To this end, we utilise mixed-effect modelling to simultaneously account for the fixed effects of melodic features in explaining participant performance, whilst also considering participant and item-level random effects, which should ensure that potentially misleading and spurious effects are accounted for; (v) we formally model change in number of recalled notes in each attempt. This latter point is based on the observation that Sloboda and Parker (1985) made, that across each attempt at singing back the same melody, participants tend to contribute more notes. However, they did not model this effect, which is an omission in the previous literature.

## Operationalising Similarity vs. Number of Recalled Notes

Alongside the similarity measures described in Experiment 1, Experiment 2 introduces the formal modelling of the dependent variable *number of recalled notes*. This represents the number of notes that a participant sings in a trial attempt. Note that this was manipulated in those conditions in Experiment 1 which affected the sung recall length: Exp. 1B (insertions); Exp. 1C (deletions); Exp. 1G (length mismatch). Overall, sung recall length was associated with higher scores across all similarity and accuracy measures. We only simulated changes in length that were less or equal to the length of the target melody. Hence, the result is intuitive: if you do not sing all notes in a melody, it cannot be fully correct; all notes must be present to have sung the melody perfectly. Note, however, the *accuracy* measure could still reach a perfect score of 1, despite singing fewer notes than in the target melody, so long as all the sung notes were in the stimulus.

As observed by Sloboda and Parker (1985), participants may take several attempts at the same melody before they manage to sing all the notes back: they consecutively build up, adding new notes to each attempt. This suggests that the sheer number of notes recalled in an attempt could be a main driver of overall similarity (numbered of recalled notes => overall similarity). However, as illustrated previously, it is not enough to simply recall the correct number of notes to obtain high melodic similarity: these notes must respect the melodic identity too (i.e., they are necessary, but not sufficient). Hence, *number of recalled notes* is *not* intended to be a measure of overall performance, but simply a count of the number of notes in each attempt, which is related to overall performance. Overall performance is measured by *opti3*, our melodic similarity variable. In Experiment 2, we model changes in numbered of recalled notes alongside changes in overall melodic similarity.

# Research Questions

We seek to answer three general questions. First, what makes melodies more easy or difficult to remember? To answer this question, we aim to construct statistical models of melody learning consider i) relevant experimental conditions (e.g., whether the melody was presented as part of a full audio recording or as melody-only *MIDI* version), (ii) features of melodic structure (e.g., melody length, tonality), and (iii) individual differences (e.g., musical background). Second, we seek to investigate the temporal aspects of learning and thus answer the question, “how do we learn melodies”? This concerns the time course of learning over multiple attempts and identifying the different aspects of learning (e.g., the types of errors made) that change across multiple trial attempts for the same melody. Moreover, focusing on temporal aspects allows us to investigate how the representations of melodies build up in memory and hence predict the type of mistakes that people make early and late in the learning process, and whether the type of mistake differs by level of prior musical experience. Finally, we ask “how do the number of recalled notes submitted change across attempt and relate to musical features, individual differences and changes in overall similarity”?

# Method

Experiment 2 uses the experimental design employed by Sloboda and Parker (1985), and subsequently used by others Oura & Hatano (1988) in different variants.

## Participants

31 adult participants (54.84% female) aged 21-38 (*M* = 26.43; *SD* = 4.43) from undergraduate courses in psychology and musicology at the University of Hamburg, Germany, were recruited. Participants’ musical backgrounds (0-25 years of instrumental training; *M* = 7.91; *SD* = 7.44) were assessed by a detailed questionnaire asking for their present and past musical activities. To be able to focus on memory and not singing errors in our analysis, participants underwent a screening phase. Participants were asked to sing three popular melodies (e.g., Happy Birthday, German national anthem etc.) of their choice that they believed they could sing error-free. Before entering the transcription and analysis of the test items, participants were selected on the basis of their rendition of these songs: Their intonation and rhythmic stability were judged by a professional singer and choir director with a longtime experience of working with lay singers. The main criteria that the choir director attended to was stable intonation and timing, as well as the ability to produce clear notes. The summary criterion was “would this person be able to join your choir?”, with the choir being an amateur level choir with singers not having received any formal singing instruction during their lifetime, and the choir only rehearsing once a week and singing 1-2 concerts per year with easy repertoire. No other technical tools were used for assessment. 23 participants that showed a largely stable intonation and sense of timing were selected on these grounds.

# Materials

The stimuli used are the same as presented in Experiment 1 (see Appendix C for a list of the test songs, as well an example from each song as musical notation, distribution of features etc.). We additionally describe some relevant features here, particularly those which are relevant to a real human participants’ ability to learn them. Melodies were between 9 and 21 seconds long (length = 15-48 notes). This extends beyond usual working memory limits and is also longer than what non-experts are usually able to recall with a high degree of accuracy on their first attempt, according to the literature (Oura & Hatano, 1988; Zielinska & Miklaszewski, 1992). 15 melodies were classified as major and 13 as minor by the Krumhansl (1990) algorithm.

All songs had a hit-like quality and an easily singable vocal melody, despite not being or having been overly popular in Germany. Participants were always asked whether they knew the songs and none indicated that they did. Hence, the melodies were unknown to them. The songs were sampled from different popular music styles as light pop, dance, ballad, rock, blues rock. Among the interpreting artists were Neil Sedaka, Dan Fogelberg, Richard Marx, Modern Talking and Paul Anka. Because all melodies came from popular western music form the last 60 years, they were all structured in phrases which can potentially be used for memory chunking. The stimulus melodies were often one stanza (line) from a verse or chorus of a pop song and contained several melodic phrases, often separated by longer notes or short rests. Hence, the full verse or chorus melody of the song would be longer than the excerpts used as stimulus. All melodies were taken from vocal passages. Note, that vocal melodies are thought to be easier to learn than instrumental passages due to the mimetic hypothesis (Cox, 2001). The singability of the melodies were piloted informally, but no melodies were discarded.

All songs were used as song excerpts from the original audio recording (audio melodies) and as single-line melody that was transcribed from the original recording and rendered in a MIDI Grand Piano sound. Melodies were transcribed from their tracks by a high-quality professional transcription service[[3]](#footnote-74). The transcriber’s brief was to transcribe the melodies as accurately as possible and notational choices were made to express what they heard as the intended structure. Because metrical information is not considered in any of the similarity measures, the notational choice of e.g., 9/8 vs. another 4/4 measure does not affect the results. Similarly, none of the similarity measures take absolute tempo or meter into account and therefore transcriptions at half time or double time would not affect similarity measurements.

The melodies were divided by random into two groups, A and B. To prevent serial effects and an uncontrolled interaction between version and melody, half of the participants listened to melodies from group A in the MIDI rendition and to the audio melodies from group B. The other half of the participants had group B melodies as MIDI and group A melodies as audio.

## Procedure

After having sung the three popular songs, participants were told that they would listen to short melodies which they had to sing back from memory immediately afterwards. They had the chance to listen to every melody up to six times and to sing them back every time again. After each sung recall, they were asked to rate their own performance on a 7-point scale for accuracy in comparison with the original, while disregarding minor intonation or other singing problems. They were asked to repeat listening and singing back each melody until the sung recall was perfect in their opinion. In doing so, participants that reached perfect recalls quickly were not forced to repeat them identically, which kept motivation high across trials. These data were not kept for analysis.

The specific instructions for the task were: “In the following you are going to hear a short melody that you should sing back immediately. Your recall (singing) is going to be recorded. Please also indicate afterwards on a scale from 1 to 7 how certain you are that the sung melody is identical to the original melody. 1 represents “very certain different” and 7 represents “very certain identical”. Please indicate also how well you knew the melody prior to this study and tell title and performer if possible”. Consequently, participants did not have to start the melody from the beginning.

Participants were first trained with two melodies, where each could be repeated up to six times. After the training phase, participants were tested with seven single-line *MIDI* melodies in the first test block. Subsequently, they were played a real song excerpt (audio melody) for training, which was followed by a test block of seven audio melodies. Having concluded the second test session, participants filled out the questionnaire on their musical background and were then debriefed. Participants were tested individually, listening to the melodies on a pair of *Beyerdynamic DTX800* headphones. Their sung recalls were recorded directly to hard disk using a *Philips MD 650* microphone and *Cool Edit Pro 1.2* as recording software device. The entire experimental session lasted about 75 minutes.

## Audio Transcription

As a result of the test sessions, approximately 2,250 audio files were obtained. For computational analyses, such audio must be transcribed to a symbolic format such as MIDI. We used the same high-quality commercial service described in Experiment 1 for the transcription of the sung recalls. To avoid any bias in the transcription process, the human transcribers were not informed about the aims and the details of the study, but a set of guidelines was provided to help with ambiguous cases (e.g., pitch bend, non-pitched sounds, rhythmic precision, and implied metrical structure). The original melodies were transcribed according to the same guidelines by the same person. However, they were not given any information about the original melodies at the time of transcribing the sung recalls in order not to introduce any bias towards the target.

## Data Transformation

After transcription, the MIDI files were converted to a tabular text format using the conversion tool *MELCONV* (Frieler, 2018) which builds on the freely available MIDIJDK library. After conversion, pitches were represented as MIDI numbers and onset times and durations were represented in MIDI beats and ticks as well as milliseconds. Time signature information was also read out from the MIDI files for later use.

# Data Analysis

## Dimension reduction of demographic variables

The questionnaire (see Appendix D1) about musical experience produced a set of mixed type (i.e., continuous, dichotomous, and polytomous) variables. To aggregate the data, we computed a pairwise correlation matrix using the mixedCor function from the *R* package *psych* (v 2.2.5) using pairwise complete observations to handle missing data (0.007% missing) and otherwise default settings. This correlation matrix was then used as the basis for factor analysis. A single-factor solution (see Appendix D2) was achieved using the cfa function from the *R* package lavaan, version *0.6-9* (Rosseel, 2012). We extracted scores using the regression method and took this variable to represent “musical experience”. The variables *age*, *sex* and *edulevel* (level of education achieved) from the questionnaire were also used as single indicator variables in the subsequent analyses.

## Main analyses

### Assessment of change in number of recalled notes and similarity scores across repeated attempts.

To begin our analyses, we inspected our descriptive empirical results. First, we assessed the mean change in number of recalled notes across successive attempts. Next, we we assessed the mean change in similarity scores (*opti3*) across attempt, as well as for the mean change in each of the individual constituent similarity measures (*ngrukkon*, *rhythfuzz*, *harmcore*; see Appendix A) across attempt.

### Correspondence between number of recalled notes and melodic similarity (opti3).

Then changes in the number of recalled notes and in melodic similarity (*opti3*) across attempt were plotted alongside each other on the same graph for comparison. For formal modelling of both number of recalled notes and *opti3*, we proceeded in a mixed effects framework. We constructed two separate models with a) *number of recalled notes* or b) *opti3* as dependent variables. Consequently, the two models assessed the development of a) the number of recalled notes (i.e. sung) across repeated attempts and b) overall improvement in performance, as indicated by melodic similarity. In our mixed effects models, participant and melody item were always included as random effects intercepts. Number of attempts and condition (*MIDI* vs. audio) were always included as fixed effects.

### Melodic feature modelling.

Subsequently, we evaluated a second set of models which additionally included melodic features as predictors. The melodic feature predictors employed were taken from the *FANTASTIC* toolbox (Müllensiefen, 2009) and are presented in Appendix E. *A priori* we chose *i.entropy* (to indicate the amount of “surprise” in intervallic information), *d.entropy* (to indicate the amount of “surprise” in rhythmic information), *tonalness* (to indicate the level of tonality), *N* (target melody length to indicate overall constraint on working memory) and *step.cont.loc.var* (to indicate the amount of variation in contour) due to previous research indicating that they serve as good predictors of melodic memory, at least in studies using a melodic recognition task (Dreyfus, Crawford, Müllensiefen, & Baker, 2016; Harrison et al., 2017; Müllensiefen & Halpern, 2014; Silas, Müllensiefen, & Kopiez, 2023). Additionally, to capture the re-occurrence of melodic patterns and the overall self-similarity of each melody (Deutsch, 1980), we compute the mean information content of each sequence of melodic pitches using the *ppm* *R* package (Harrison, Bianco, Chait, & Pearce, 2020). After iteratively eliminating predictors with a non-significant main effect contribution, we tested the interaction between significant feature predictors and attempt. Our feature-based modelling approach is closely related to Baker (2019)’s modelling of melodic encoding and recall processes used in melodic dictation among musicians who also makes use of features computed from the *FANTASTIC* (Müllensiefen, 2009) toolbox.

### Individual differences modelling.

Since Sloboda and Parker (1985) dichotomised their participants into “non-musicians” and “musicians”, as a means of comparison with their data, we produced some figures as change in dependent variables across attempt but stratified into two groups: high musical experience and low musical experience. These groups were derived by taking the median value of musical experience variable and grouping into two bins based on this. Those with a musical experience value below or equal to the median were classified as being in the lower musical experience group, whereas the rest, the higher musical experience group. There were 12 participants in the former and 11 in the latter.

Subsequently, we extracted the random effects intercepts for each participant from each of the two (number of recalled notes vs. overall similarity) models. We took these values to represent a participant-level latent score reflecting a) the number of recalled notes they can hold in memory b) their overall melodic recall ability. To evaluate whether musical experience is a good predictor of individual differences in both number of recalled notes and overall melodic recall performance, we regressed these participant intercepts onto the participant musical experience score derived earlier. The incremental modelling approach described in the above steps broadly follow the suggestions of Long (2011).

### Mediation analysis.

As a means of formally associating number of recalled notes and *opti3* with one another, as well as to connect melodic features to *opti3*, we computed a mediation analysis whereby melodic features acted as predictors, number of recalled notes acted as mediator, and *opti3* as dependent variable.

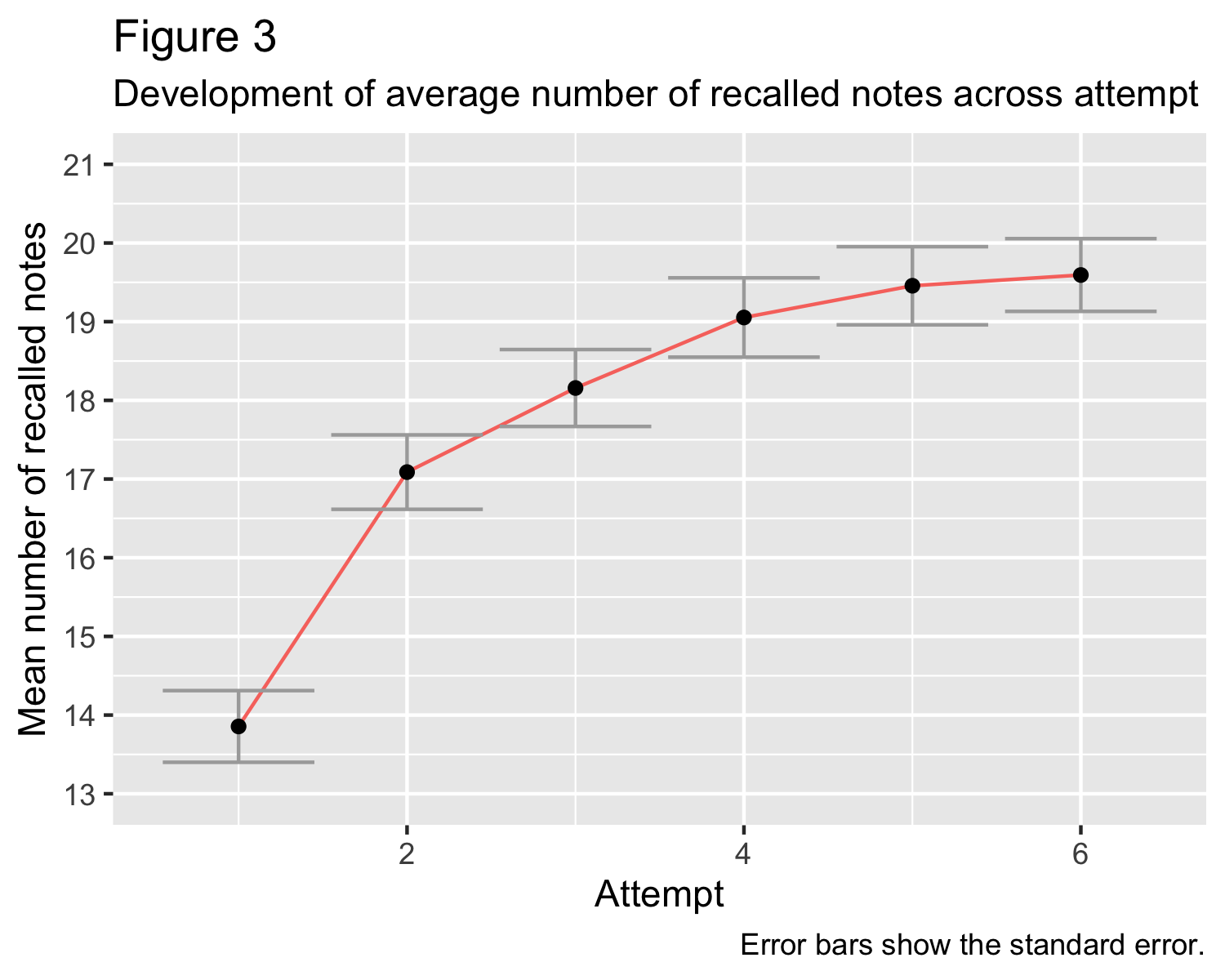
### Correspondence between number of recalled notes and melodic similarity (opti3): revisited.

Lastly, we revisit the association between number of recalled notes and melodic similarity but using accuracy-style measures. Specifically, we view and model the number of recalled notes as a changing proportion of correct and incorrect notes across attempts.

# Results

## Assessment of change in number of recalled notes across attempt

To visualise the change in number of recalled notes submitted across trials, Figure 3 graphs the mean number of recalled notes, as a function of attempt. As shown, the number of recalled notes increases across successive attempts. The effect is clearly non-linear, with a diminishing gain in number of recalled notes per attempt. Note that the average melody length is *N* = 25.39. Consequently, even after six attempts, on average, participants are still not submitting close to the number of notes in a target melody.[[4]](#footnote-89)



In the formal mixed effects model where number of recalled notes was dependent variable (Model *A1*), the estimates of the fixed effects coefficients were *B* = 3.53 (*p* < .001) for log attempt and *B* = 5.18 for condition (*p* = .02). The latter result suggests that hearing a melody as a full audio excerpt is associated with five more notes being recalled to an attempt on average. The marginal R2 value of the mixed effects model was 0.14 and the conditional R2 value 0.65 (Nakagawa & Schielzeth, 2013). This suggests that the fixed effects (attempt and condition), whilst significant, explain a relatively small amount of variance compared to the random effects (melody, and participant). Adding an interaction term for the random effects interaction between participant and melody considerably increases the conditional R2 value (to 0.79) and the marginal R2 value slightly (to .143). This final model (Model *A1.2*) is shown in Table 5. The increase in number of recalled units across attempts seen here corresponds to previously developed computational models of singing production and general serial recall (Anderson, 1972; Chikhaoui et al., 2009).

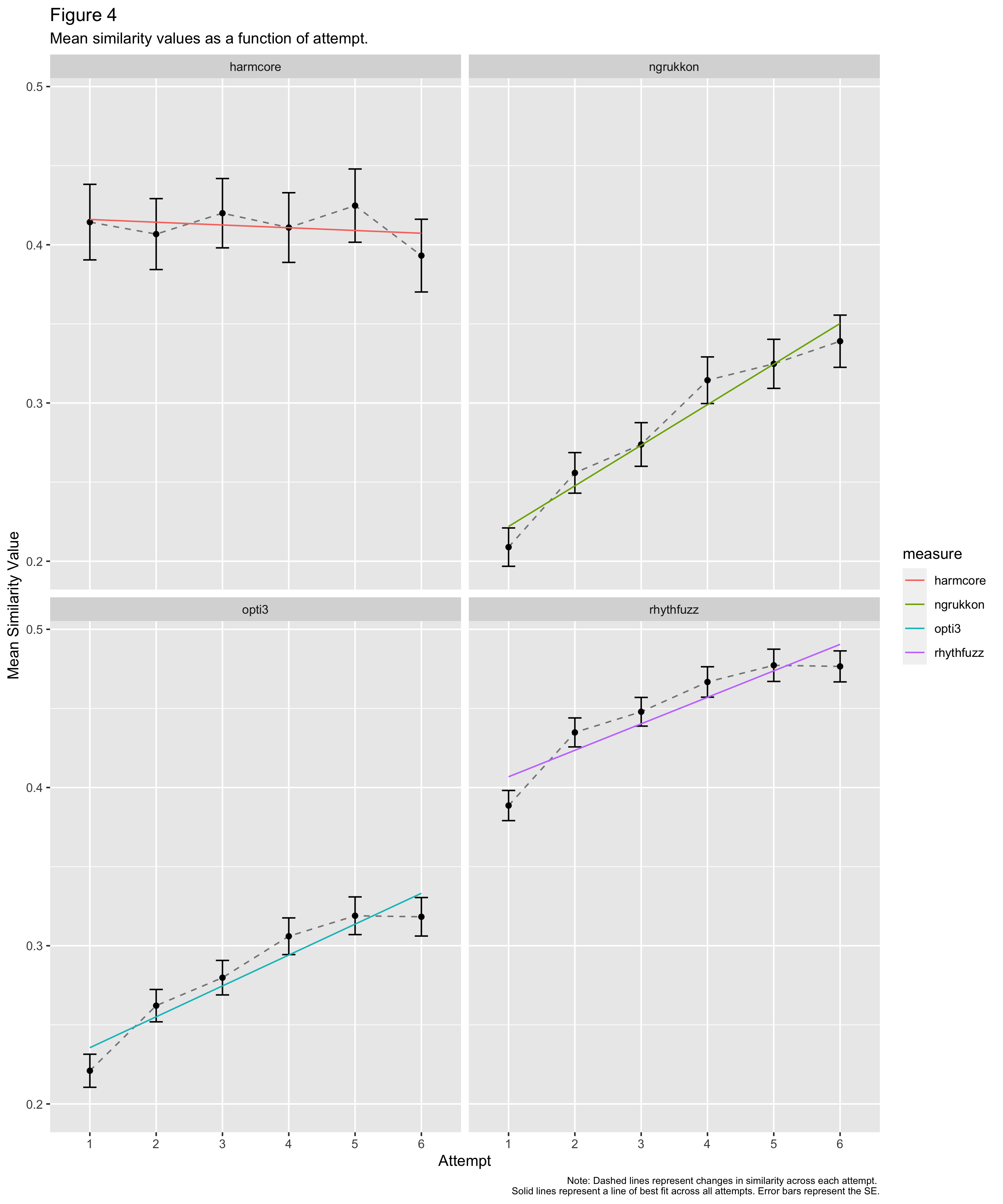
Table 5:

*Model A1.2: Mixed effects model regressing number of recalled notes onto attempt and condition. The index ‘S’ refers to the ‘Sound’ condition.*

| Term |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 11.49 | [8.25, 14.73] | 6.95 | 32.20 | < .001 |
| ConditionS | 5.26 | [0.95, 9.57] | 2.39 | 25.66 | .024 |
| Logattempt | 3.65 | [3.32, 3.98] | 21.89 | 1,466.34 | < .001 |

## Assessment of similarity scores across repeated attempts

To visualise higher level changes in melodic recall performance (indicated by similarity) across attempts, Figure 4 graphs the mean score of each similarity measure (*opti3*, *ngrukkon*, *harmcore* and *rhythfuzz*), as a function of attempt. A linear model (represented by solid-coloured lines) would suggest a general increase across attempt for all variables, except *harmcore*, which appears relatively stable across attempt (see Appendix G for linear model details). However, as seen with number of recalled notes, whilst a linear model predicts the data over the course of six trials reasonably well, for the generally increasing variables (*opti3*, *ngrukkon* and *rhythfuzz*) a non-linear effect (i.e., with diminishing gains across attempt) seems to represent the data better.



An equivalent to model A1 was fitted using *opti3* as dependent variable (Model B1). Both predictors were significant in the model: log attempt, = .07, *p* <. 001; condition, = .10 (*p* = .01) The latter suggests that hearing a melody in its full audio is associated with a .10 increase in similarity of recall to target melody, as indicated by *opti3*. The model achieved a marginal R2 of .098 and a conditional R2 of .49, again suggesting that fixed effects explain a relatively small amount of variance compared to random effects (melody item and participant). Adding the interaction term between melody item and participant random effects again considerably increased both the marginal R2 (to .101) and the conditional R2 (to .71). See Table 6 for the final model (Model B1.2).

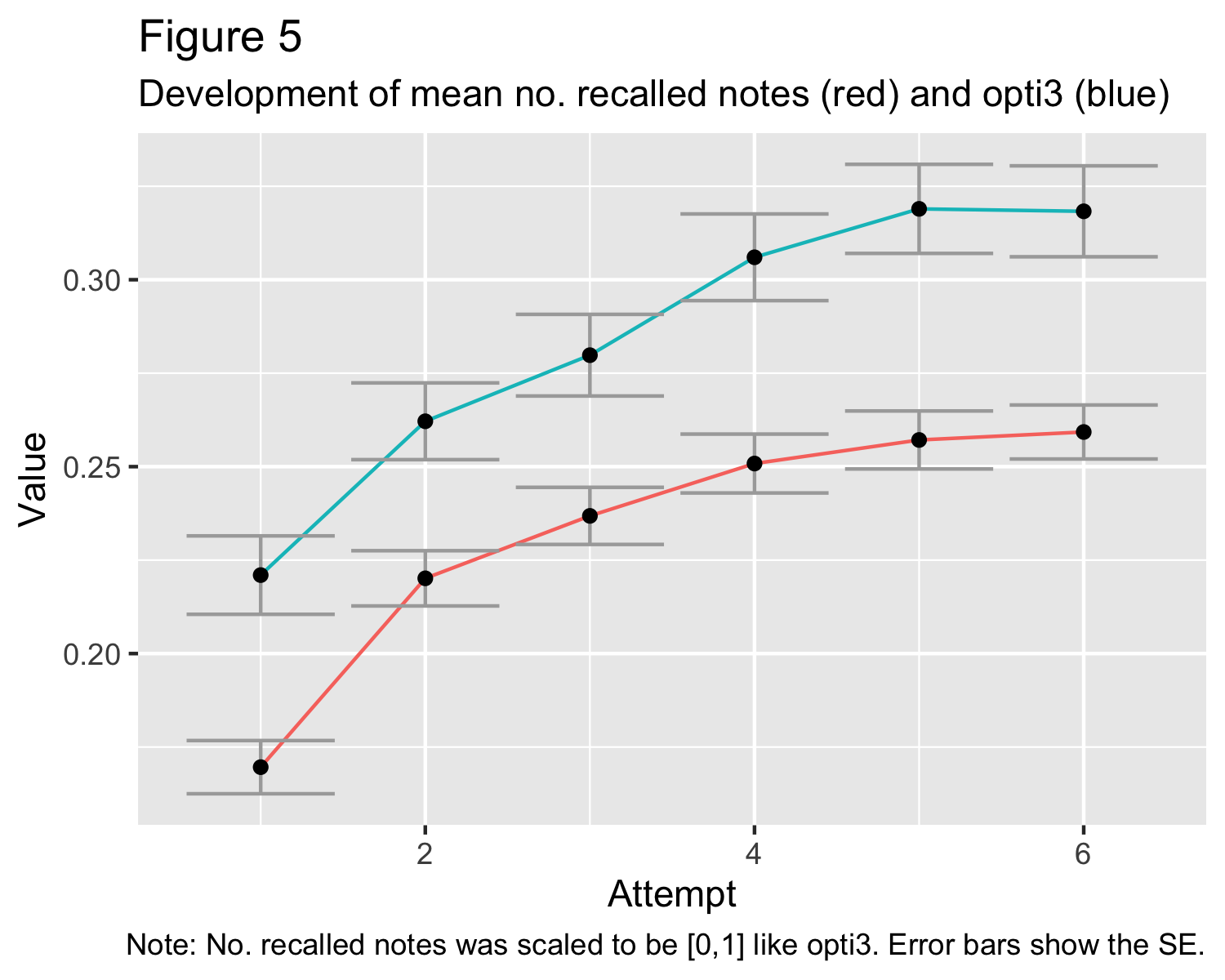
Table 6:

*Model B1.2: Mixed effects model regressing the similarity of melodic recalls (opti3) onto attempt and condition.*

| Term |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 0.16 | [0.10, 0.23] | 5.00 | 37.76 | < .001 |
| ConditionS | 0.10 | [0.02, 0.18] | 2.49 | 25.54 | .019 |
| Logattempt | 0.07 | [0.06, 0.08] | 15.88 | 1,460.86 | < .001 |

# Correspondence between number of recalled notes and melodic similarity (opti3)

The development of both overall similarity (*opti3*) and the number of recalled notes across attempt is broadly similar in shape: increasing, with diminishing gains on each attempt. To make this clear, we rescale the number of recalled notes variable to be in the range 0 to 1, like *opti3* and plot them alongside each other in Figure 5.



This convergence is interesting, and we hence suggest that, in order obtain a comprehensive picture of the cognitive processes involved in melodic recall, it is necessary to model the number of recalled notes alongside overall change in melodic recall performance (here indicated by melodic similarity) across recall attempts. Consequently, we proceed by modelling the two effects via two sets mixed effects models in parallel. We take forward both models *A1.2* and *B2.2* (Tables 5 and 6), where i) condition and log attempt are always included as fixed effects and ii) participant and melody item, plus the interaction between participant and melody item, are random effects, as the basis for the remaining analyses.

Note that, it is not only important to understand the degree to which number of recalled notes and *opti3* converge, but diverge, and hence, measure different constructs. The bivariate correlation between the two is *r* = .42 suggesting that, as expected, they are related to a moderate degree, since as we noted earlier, *opti3* is dependent on the length of comparison targets in a “soft” sense. However, since the correlation is only moderate, it confirms empirically, and with human participant data, that *opti3* measures something beyond the length of comparison targets (i.e., the harmonic, rhythmic and intervallic information it is intended to capture).

# Melodic feature modelling

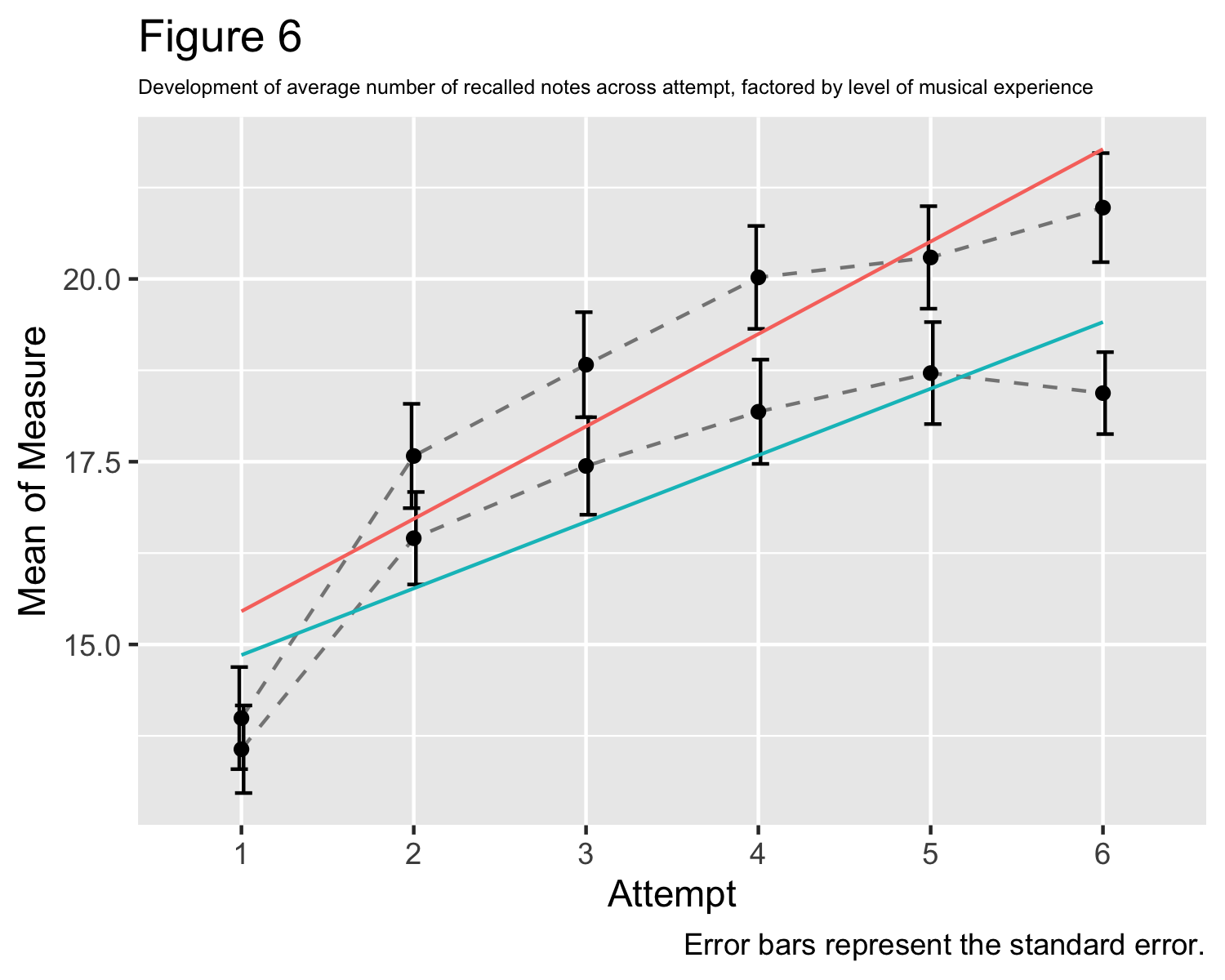
For modeling the memorability of melodies we added *N* (melody length), *tonalness*, *i.entropy*, *step.cont.loc.var*, *d.entropy* and *mean\_information\_content* as additional predictors to the mixed effect models described before. With all predictors in, the marginal R2 increased significantly from 0.14 to 0.46 when *number of recalled notes* was dependent variable and marginally from 0.10 to 0.16 when *opti3* was dependent variable. However, after removing non-significant predictors, only *N* was a significant predictor when *number of recalled notes* was dependent variable, and none were significant with *opti3* as dependent variable. In a model with *number of recalled notes* as dependent variable, and only *N* as fixed effect predictor alongside log attempt and condition, the marginal R2 was 0.44, suggesting that the other melodic feature predictors really do not add much explanatory power to the model. In Appendix H, we also present variance inflation factors and partial R2 values for diagnostics. Altogether, the interpretation that the non-significance of the other melodic features is due to high collinearity can be ruled out, and it is evident that melody length (*N*) substantially explains variance in *number of recalled note events* by itself. Note also, as shown in Appendix C2, several melodic features have as much variance as melody length, as indicated by higher coefficient of variations, which facilitate the comparison of the SD across measures. This also suggests that melodic features have enough heterogeneity beyond melody length, which empirically has *less* heterogeneity.

We tested the interaction between *attempt* and *N* in the model with *number of recalled notes* as dependent variable. The interaction term was statistically significant ( = 0.21, *p* < .001), suggesting that the length of the melody differentially affects the number of recalled notes, depending on the attempt number. For this final model (A2.2), the marginal R2 value was 0.45 and the conditional R2 value 0.80.

## Individual differences modelling

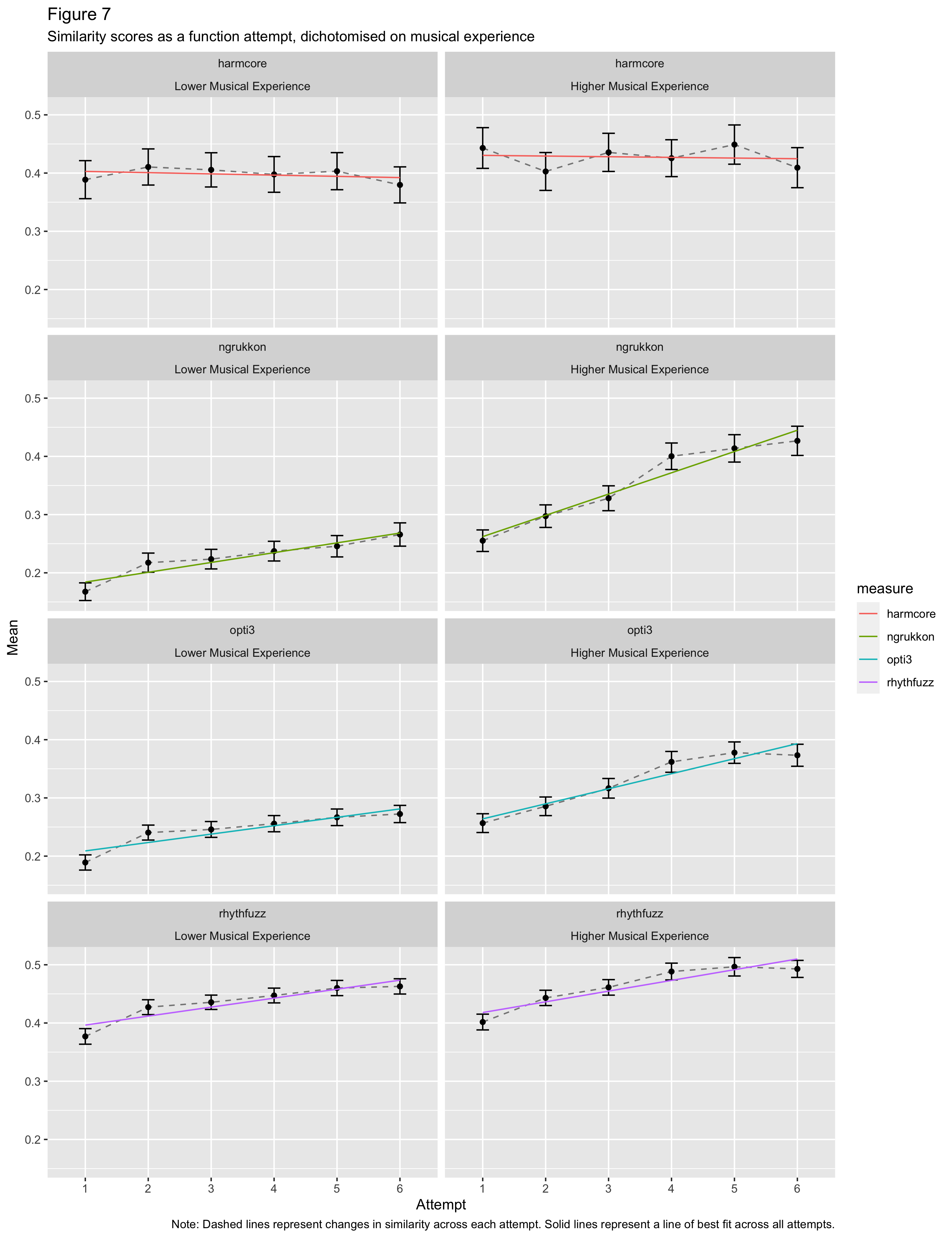
### Individual differences in changes of number of recalled notes across attempt.

Figure 6 presents changes in number of recalled notes across attempt, based on the median split on musical experience described earlier. Broadly speaking, the pattern of results in low and high musical experience groups is similar, with a non-linear increase in number of recalled notes across attempt for both groups. However, whilst both groups submit a similar number of recalled notes to the first attempt on average, the higher musical experience group tend to submit more notes to each subsequent attempt (however, see Appendix F1 for a by-participant comparison). This is most notable in attempt two, where there is a larger increase in notes for higher musically experienced participants than for lower musically experienced participants. A linear model across all trials does not seem to describe the data well, but is useful for comparing the general slopes, which appear to be approximately the same, though the higher musically experienced group’s slope appears somewhat steeper. This suggests that higher musically experienced participants may be able to learn more quickly by extracting more melody notes in memory on each successive attempt compared to lower musically experienced participants. We do not model this artificial dichotomisation of musical experience formally.



### Individual differences in similarity changes across attempt.

The corresponding figure for similarity (*opti3*), Figure 7, suggests that the higher musical experience group generally have better melodic recall, as indicated by generally higher similarity scores across trials (i.e., a larger intercept). The slopes (i.e., rate of increase) across attempts appears to be approximately similar, except for with the *ngrukkon* measure, which suggests that, across successive attempts, those with more musical experience improve the interval accuracy of their recalls more effectively than participants with lower musical experience. The difference in slopes is also somewhat notable for overall similarity (*opti3*). See an alternative by-participant visualisation in Appendix F2.

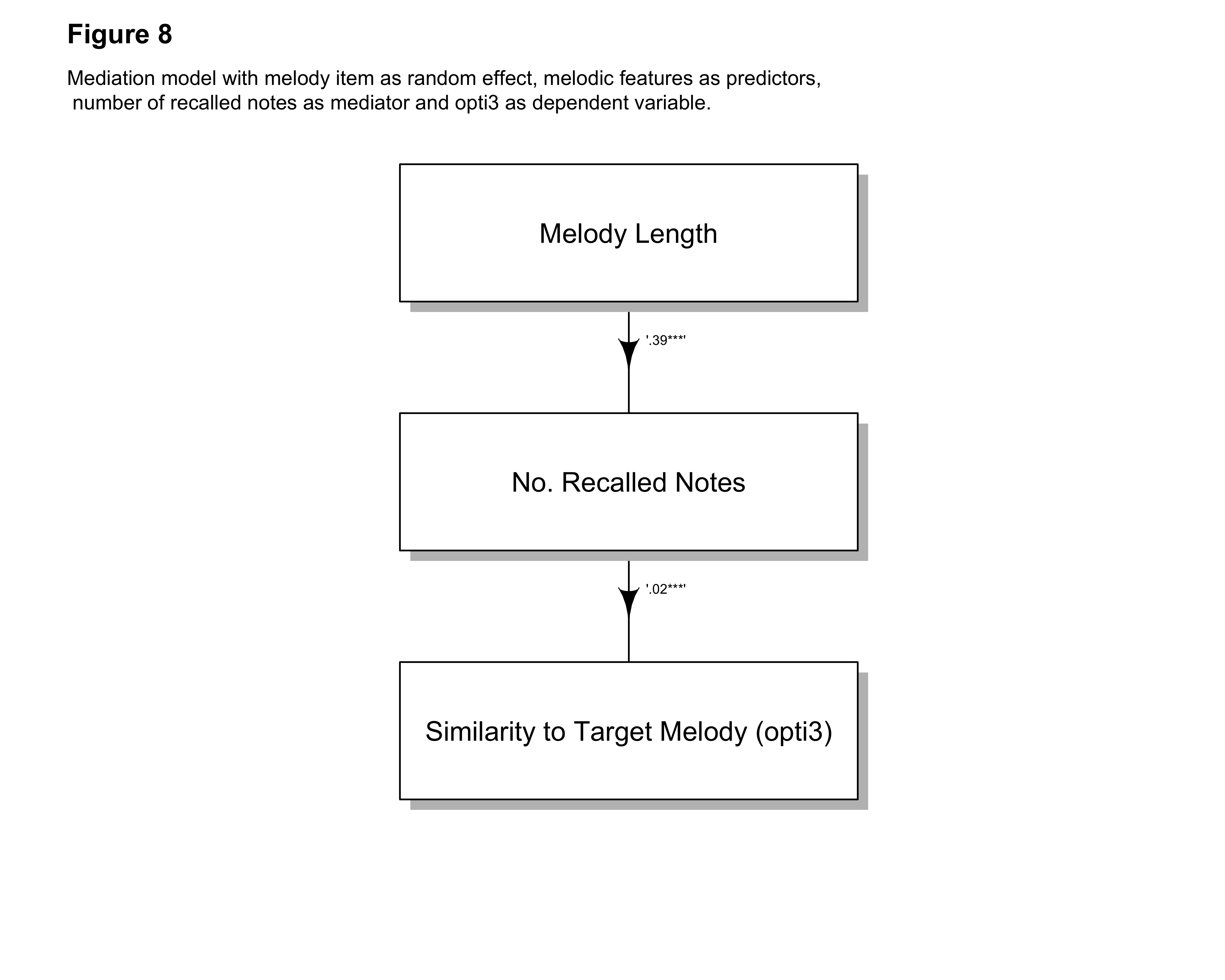


To model and explain some of the random effects variance attributable to participant, random effect intercepts were extracted for each participant, from each of the two most-developed models described earlier (A2.2, B1.2). These were taken to represent two participant-level latent melodic recall processes: one specifically to do with abilities concerning the number of recalled notes, and the other with overall level of melodic recall (as indicated by the *opti3* measure of similarity). When regressing the participant random intercepts from the *number of recalled notes* model onto *musical experience*, *age*, *edulevel* and *sex* in a general linear model, only musical experience was a significant predictor. Removing the other variables left a model with a moderate R2 value of 0.36 (adjusted = 0.33), *p* < .01, and musical experience as the sole significant predictor, = 0.05, *p* < .01. A similar pattern was seen for the model built with *opti3* as dependent variable: only *musical experience* was a significant predictor ( = 1.63, *p* < .01). This model had a small R2 value of 0.24 (adjusted = 0.20).

## Mediation analysis

Earlier, when adding melodic features as fixed effects to the base mixed effects models described above, we found that only *N* was a statistically significant predictor of performance when *number of recalled notes* was dependent variable and none when *opti3* was dependent variable. However, as noted previously, there is a correspondence between number of recalled notes and overall similarity across attempts (see Figure 5). Perhaps melody length (*N*) could indeed predict overall performance (*opti3*), via an effect on number of recalled notes. This hypothesis can be implemented as a mediation analysis: melody length predicts performance, via the number of recalled notes submitted to the trial. We computed this hypothesis using the *mediate* function from the *R* package *mediation* (*v* 4.5.0). *N* was the single predictor, *opti3* was the dependent variable and *number of recalled notes* was the mediator. Unfortunately, it is not possible to specify more than one random effect with this functionality, so two separate models were computed, one with melody item and another with participant as random effects.

In the model with melody item as random effect, the Average Direct Effect = .07 [-0.005, 0.15], *p* = .07 was not statistically significant. However, the Total Effect = 0.11 [0.03, 0.20], *p* = .001 and the Average Causal Mediation Effect = .04 [0.01, 0.07], *p* = .01 were statistically significant. This suggests that longer melodies lead to more note events being recalled (because the target melody is longer, so requires more notes), which in turn increases the *opti3* score. See Figure 8 for a representation of this model.



In the model with participant as random effect, the Average Direct Effect = .05 [0.002, 0.09], *p* = .04, the Total Effect = 0.10 [0.05, 0.14], *p* < .001 and Average Causal Mediation Effect = .05 [0.03, 0.07], *p* < .001 were statistically significant. This is a similar pattern as the last model, with similar coefficients, implying the same interpretation we gave earlier. However, we suspect that the Average Direct Effect being statistically significant here is spurious and would vanish were we able to model melody item as a random effect simultaneously. We do not represent this second model as a figure because its interpretation is then similar to that presented in Figure 8.

That the Average Causal Mediation Effect was statistically significant in both models suggests that melody length can indeed be a predictor of overall performance (as indicated by *opti3*), at least partly via its influence on number of recalled notes submitted to an attempt. When comparing nested models with *N* as fixed effect predictor of *opti3*, with and without number of recalled notes, the additional variance explained when including number of recalled notes was justified: in the case of the model with melody item as random effect, the model with *number of recalled notes* had a lower *BIC* value (-1614) than the one with (*BIC* = -1310); the same was true for the model with participant as random effect (*BIC* = -1322 and *BIC* = -1075 respectively). See Tables 7 and 8, respectively.

Table 7. Model comparison with and without number of recalled notes as fixed effects predictor. fit.dv = with number of recalled notes; fit.totaleffect = without number of recalled notes

|  | npar | AIC | BIC | logLik | Chisq | Df | Pr(>Chisq) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **fit.totaleffect** | 7 | -1310 | -1271 | 661.8 |  |  |  |
| **fit.dv** | 8 | -1614 | -1571 | 815.2 | 307 | 1 | p < .001 |

Table 8. Model comparison with and without number of recalled notes as fixed effects predictor. fit.dv2 = with number of recalled notes; fit.totaleffect2 = without number of recalled notes.

|  | npar | AIC | BIC | logLik | Chisq | Df | Pr(>Chisq) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **fit.totaleffect2** | 7 | -1075 | -1037 | 544.7 |  |  |  |
| **fit.dv2** | 8 | -1322 | -1278 | 668.9 | 248 | 1 | p < .001 |

# Modelling the correspondence between *opti3* and *number of recalled notes* revisited

Sloboda and Parker (1985) observed that the number of errors are relatively stable across attempts, whereas the number of correct notes increase in a non-linear fashion. The increase in number of correct notes appears to correspond to the increase in number of recalled notes that are being submitted per attempt, which too increases non-linearly across attempt. However, this was not formally modeled in their original study. With our data, Figure 9 visualises the number of recalled notes per trial as a proportion of correct and incorrect notes, along with *opti3*, rescaled to be on the same scale as number of recalled notes. Figure 9 highlights two general patterns: i) the non-linear increases of number of recalled notes and similarity (overall performance) across successive attempts, alongside one another; ii) the stability of number of errors.

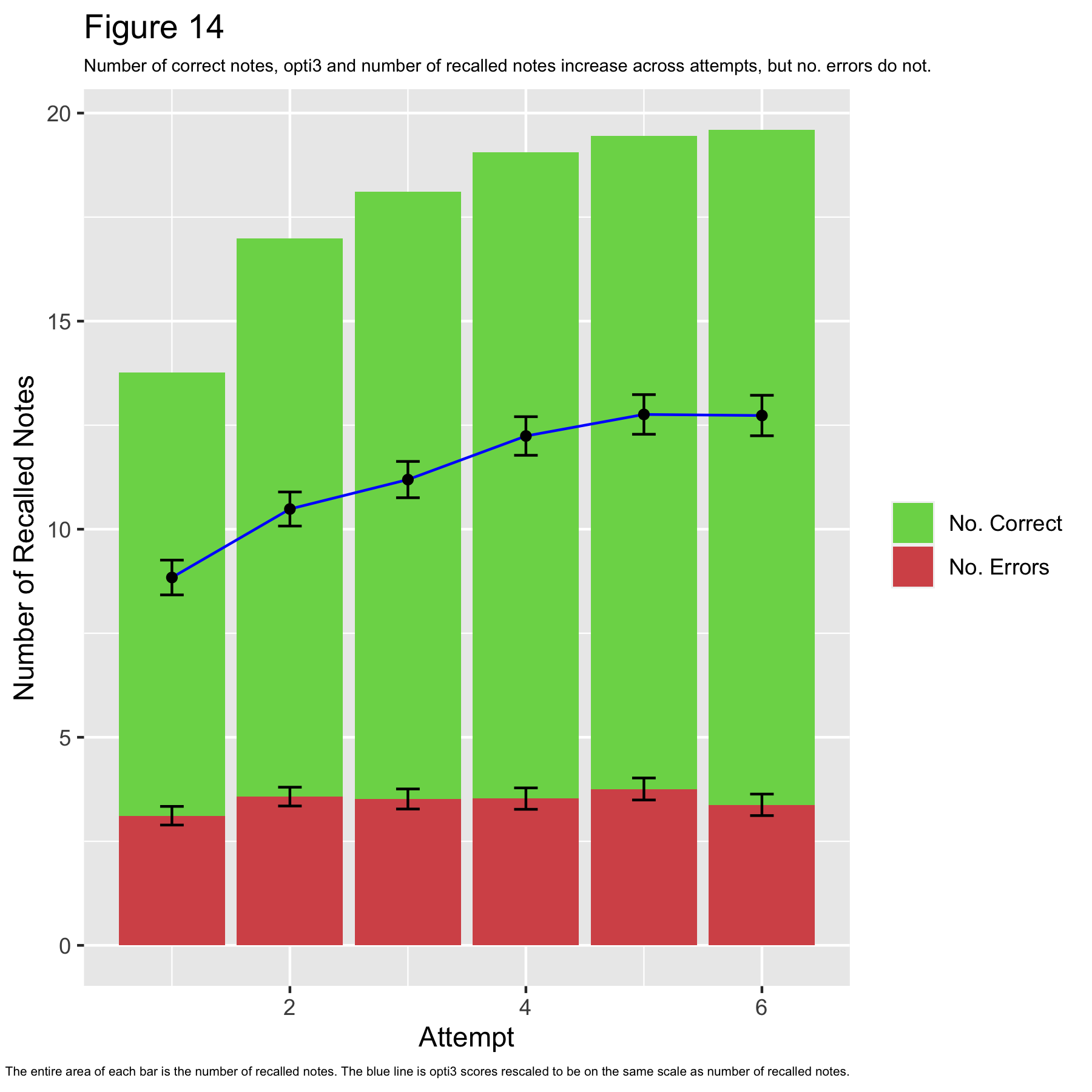


Figure 9 suggests that the number of recalled notes to each attempt could, in turn, be a predominant factor in determining the overall improvement in similarity across each attempt. Specifically: i) across attempts, participants submit more notes on each new attempt; ii) The number of errors stays approximately the same across attempt, but the overall number of correct notes go up across attempts. In other words: more of the new recalled notes are correct; iii) Consequently, similarity goes up over attempts. We have already modelled effects i) and iii) (Tables 5 and 6). To model ii) we fit a two separate linear regression models where attempt is a predictor of either correct notes or errors. As shown descriptively in Figure 9, in the model where no. correct notes is dependent variable, attempt is a significant predictor (; *p* < .001). However, when no. errors is dependent variable, attempt is non-significant with a near-zero -coefficient (; *p* =.27).

# Discussion

In Experiment 2, we sought to study melodic learning using the melodic recall paradigm. First, we were interested in assessing how melodic recall changes across multiple attempts. Second, we aimed to assess whether certain musical features could predict performance on the task. Third, we contended that modelling the change in number of recalled notes submitted to a trial was important and noted that such modelling was omitted from previous melodic recall research (Ogawa et al., 1995; Sloboda, 1985; Zielinska & Miklaszewski, 1992). Other literature describing non-musical verbal recall (e.g., Anderson, 1972) has already described some relevant models which serve as non-musical analogues to the observations Sloboda and Parker (1985) made, that the number of recalled notes increase across attempts. Taking inspiration from the serial recall literature (Anderson, 1972; Chikhaoui et al., 2009), we modeled the change in number of recalled notes across attempt, alongside changes in overall melodic similarity. We investigated how both relate to individual differences and melodic features and, in the end, suggested how the variables can be combined in an integrative fashion via mediation analysis.

## How do we learn melodies?

To understand how we learn melodies, we studied how both the number of recalled notes and overall similarity change across the time course of six attempts. The number of recalled notes starts from an incomplete recall in attempt one, with each subsequent attempt adding more notes than the previous. This grows as an exponential curve which asymptotes at the number of notes in a melody, with six attempts potentially not being enough for all melodies. These results are similar to those presented in the (non-musical) free recall literature, where the learning curve approximates an exponential curve with an asymptote equal to the number of items in a target list (Anderson, 1972; Murdock Jr., 1960). The exact shape of this curve depends primarily on the number of notes in the target melody and the participant. Generally, similarity between the target melody and the sung recall increases across attempts too, suggesting incremental learning across repetitions. This is the case for overall composite melodic similarity (*opti3*) and the constituent parts of that: rhythmic similarity (*rhythfuzz*) and note similarity (*ngrukkon*). However, harmonic similarity (*harmcore*) does not change across attempt. This has previously been interpreted as suggesting that tonality is extracted earlier in attempts (Sloboda & Parker, 1985), whereas other features leave more room for improvement.

In general, we argue that the patterns in our data are not sufficient to argue that certain musical features are extracted earlier or more readily than others. This may strike the reader as curious, since our data shows similar patterns to those reported previously by Sloboda and Parker (1985) (i.e., harmonic learning is stable relative to other domains, such as rhythm and intervallic structure, which clearly increase across attempts). The difference is in our interpretation. We suggest that, simply because harmonic learning does not increase across attempts, it does not prove that harmony is extracted earlier or more readily in memory, only that it does not increase across attempts, for some reason currently unknown. Whilst the melodies we and other previous studies used as stimuli may be in different keys to each other, *within* each melody there tends to only be a single key (i.e., there are no modulations). However, melodies may visit different related modalities within a single key (e.g., chord I going to its dominant). In this sense, harmony and tonality are not “manipulated” to the same degree as intervals and rhythm. But, our approach was to use melodies from real popular music songs, without artificial manipulation. Consequently, this simply implies that, since Western pop music generally contains melodies in a single key, they naturally contain little tonal variance, whereas intervallic and rhythmic structure naturally have more variance. Hence, without manipulating the harmonic structure of melodies to contain more tonal variance (changing key/tonal centre), we can only comment on the statistical regularities of melodies that arise in popular music, and their incidental association with memorability. To establish whether tonality is extracted more easily than other features, beyond its natural variance, further research would need to use melodies which more clearly change key. Then, we suspect that we would then observe clear improvements across multiple attempts in the harmonic domain too.

Instead of representations for musical features developing in memory separately, we suggest that representations for melodies may build up simultaneously across domains. Not only would this be more cognitively efficient, but it would be in line with the tendency for different melodic features to correlate with one another (Baker, 2019). In other words, if different features correlate with one another (e.g., phrase endings contain both longer notes and more salient scale degrees, like the tonic), the brain should implicitly extract such co-occurring statistical regularities (Pearce, 2018). Whether or not this particular interpretation is true, we highlight our finding that improvement *can* be indicated across trials, which *not* in accordance with Sloboda and Parker (1985), who only observed that attempts got longer (i.e., more note events were recalled), but not better. That both number of recalled notes and performance increase across trials has been observed in other domains too (e.g., see Koh, 2002).

With respect to individual differences, firstly, in our mixed effects model we found an interaction between the random effect of participant and the random effect of melody. This suggests that certain melodies are more readily remembered or learned by certain participants than others. Broadly, this could be because some participants have previously implicitly learned similar melodies to those they were tested with. Alternatively, perhaps some melodies contain features which rely more on musical vs. nonmusical memory than others: the former might benefit highly musically experienced participants, and the latter, those with very good non-musical memory, but not necessarily very good music-specific memory. In this way, we also observed how, generally, participants with higher musical experience seem to perform better and demonstrate steeper learning slopes, suggesting that they learn melodies more quickly on average, indicated also by musical experience being a significant predictor of both number of recalled notes and overall similarity. However, this is not the case with all participants (see Appendix F1). Some participants low on musical experience can still learn quickly across trials presumably because, even though they have low levels of musical experience, they can nonetheless make large improvements over trials, probably as a function of other abilities, such as their general working memory. As explored in Silas et al. (2022), high general working memory may *predispose* people towards music training, explaining the general finding that musicians/those with more music training tend to have higher general working memory abilities (Talamini et al., 2017, 2016). Thus, at least some of the variance which explains musicians’ superior *musical* abilities is attributable to their already very good general working memory. The role of general working memory probably also explains why the number of recalled notes increases across attempts, reflecting general capacity constraints (Cowan, 2010), rather than music-dependent memory.

In our data, and as Sloboda and Parker (1985) previously noted, it seemed that a dominant factor in performance is the number of recalled notes submitted to an attempt. This effect might be more to do with sheer rote repetition (i.e., multiple attempts), rather than extracting musical features. In this way, each attempt is a new iteration adding more note events to the long-term memory store for a particular melody. Whilst we would expect the extraction of musical features to somewhat mitigate general capacity limits (because structure helps memorability Gobet et al. (2001); Gobet (2005); Thalmann, Souza, and Oberauer (2019); Müllensiefen and Halpern (2014)), perhaps it is not so surprising that the sheer number of recalled notes might be so influential. With respect to general theories of working memory capacity constraints, the lengths of the melodies we used were relatively long (*N* = 15-48, *M* = 25.39, *SD* = 8.67). Even though in the real world melodies tend to be longer than this, our melodies reflect good ecological validity (being taken from commercial pop songs), and other melodic features are relevant to memory, it still seems reasonable to suggest that central capacity limits in working memory could be broadly responsible for the producing the constraints on number of recalled notes produced in an attempt, as we and Sloboda and Parker (1985) observed (Cowan, 2010; Miller, 1956; Shiffrin & Nosofsky, 1994; Vergauwe, Barrouillet, & Camos, 2010). This is at least the case for when melodies are long enough (e.g., 15-48 notes) to require multiple attempts to sing back in full. In other sung recall research with short unknown melodies 3-15 notes in length (Silas, Müllensiefen, et al., 2023; Silas, Robinson, et al., 2023), where it is conceivable to sing back a melody in one attempt, we have been able to successfully connect melodic features to melodic similarity (*opti3*) directly. This suggests that, particularly with longer melodies, general memory capacities are important to include in modelling, beyond musical features and musical memory (Silas et al., 2022).

## What makes melodies difficult to remember?

Consequently, when including variables to indicate melodic complexity as predictors in our mixed effects model, we found that they were not significant predictors of melodic recall performance with *opti3* as dependent variable. As noted, however, the number of recalled notes seemed to be a main factor that might be dominating overall melodic recall performance - and melody length (*N*) was a significant predictor of number of recalled notes. That our measure of similarity, *opti3*, also across trials indicated that similarity might increase predominantly as a function of the number of recalled notes. Taking this observation into account and including *number of recalled notes* as a mediator between melody length (*N*) and *opti3* as dependent variable, connected melody length to *opti3*, as indicated by the average indirect causal effect being statistically significant. This suggested that the length of the target melody predicts melodic recall performance via the sheer number of recalled notes submitted to a trial attempt. Specifically, longer melodies tend to lead to more notes being recalled (since they are longer, and require more notes); more notes being recalled tends to increase overall similarity, but longer melodies are also more difficult to recall.

Non-musical models of serial recall predict similar effects to those seen in our data. For example, as Anderson (1972) notes, Murdock Jr. (1960) “concluded that the free recall learning curve was exponential with an asymptote equal to the number of words in the list”. A similar asymptoting effect can be seen in our data, although, lower than the average number of notes in a target melody. This might suggest that, whilst musical features could make melodies more or less difficult to remember, perhaps they are secondary to the sheer length of the melody itself, and the current working memory load (Baddeley & Hitch, 1974), at least: i) with the current melody set and data and ii) when the length of melodies are long enough to require multiple attempts to remember all the notes. However, with shorter melodies, melodic features should matter more than the overall length. Consequently, when studying melodic recall, both musical features and the sheer number of recalled notes should be integratively modeled. In this paper, we did this via mediation modelling. In further research, more detailed relationships including other variables (e.g., musical training or general musical sophistication, general working memory) should be explored, using larger and more heterogenous samples of both melodies and participants.

Lastly, the experimental factor condition (audio vs. *MIDI*) was a significant predictor of performance. This suggests that when a melody is learned from its full audio, rather than symbolic representation, it is more easily learned. Presumably, this is because acoustic features help learning (Salakka et al., 2021), as well as other cues like lyrics and the human voice, which may help memory through the eliciation of verbal memory and social psychological systems (Clayton, 2008; Tarr, Launay, & Dunbar, 2014) respectively.

## Summary and Conclusions

Consequently, melodic representations build up over multiple hearings (and sung recalls). On each attempt, the main constraint appears to be the working memory load (Baddeley, 2000; Baddeley & Hitch, 1974), limited to a certain number of notes that can be recalled. Melodies with less complex features may potentially help the number of notes that can be recalled, but the main feature that determines recall is the length of the melody to be recalled, should it require multiple attempts to sing all the notes back. On each attempt, more notes will be recalled, and the number of recalled notes will approach the number of notes in the target melody [similar to models of non-musical serial recall; Anderson (1972)], or a long-term memory constraint dependent on the timespan of learning. But, so long as the error rate remains stable, and the number of recalled notes does not exceed those in the target melody, the overall similarity to the target melody will increase across attempts. Formal musical experience and training should aid memory and to help learn melodies quickly, presumably because of mental templates which help structure the melody and more efficiently integrate it into memory (Chenette, 2021). This may be similar to the notion of *long-term working memory* (Ericsson & Kintsch, 1995). However, such musical expert memory may not be necessary, but rather, sufficient: a very good general memory and little formal musical experience may help in any case. After all, perhaps singing back pop melodies is more like a general ability and essential part of human life, rather than a formally trained musical activity.

## Limitations

We suggested that number of recalled notes could be a driver of *opti3* scores. However, we note this is a logical assumption, but not fully deductive. In other words, our data cannot fully prove the causal chain that increase in number of recalled notes across attempts are causally responsible for improvement in *opti3* across attempts. For instance, it could be that *opti3* increases across attempt alongside *number of recalled notes* simply in an associational manner, whereby *opti3* increases despite the associated increases in *number of recalled notes*. However, beyond the strong associational pattern, there are strong logical and inductive grounds for supposing the causality. Most importantly, the *opti3* measure of similarity is dependent on the length of the melodies to be compared in only a soft sense, which invokes a causal mechanism. However, this does not imply that all the variance explained in *opti3* is attributable to *number of recalled notes*. Hence, we highlight to the reader that we are arguing for the plausibility of causality, rather than inferring one.

That fact that we did not counterbalance the order of MIDI/audio (i.e., all participants heard MIDI excerpts then audio excerpts) could potentially be a confounding factor and contributed to more notes being recalled in the audio condition. Perhaps people became more confident or simply better at singing back across the course of experiment. However, we suspect this might have been a small effect compared to the advantage of having lyrics as well as expressive cues and musical information form the backing track that helped participants to remember more notes form the full audio.

# Future Directions

Our study suggests several future directions for research with the melodic recall paradigm. First, we suggest that number of recalled notes and *opti3* should be even more integratively modelled, using a much larger database of items and more heterogeneity in melodic features. We have recently implemented such a framework (Silas, Müllensiefen, et al., 2023). Second, a general working memory construct (measured by one or more variables) should be included as a predictor, as this may have some explanatory power aside from musical memory faculties (see Silas et al., 2022 for a discussion of these issues). Third, new research suggests other melodic features, such as symmetry, may be interesting to explore as melodic feature predictors (Clemente et al., 2020; Herborn, 2022). Fourth, as we have argued elsewhere (Silas, Müllensiefen, et al., 2023) singing accuracy and melodic recall measures should be simultaneously measured to understand and represent both domains properly. Lastly, since effects around item length are modelled and described well in the *ACT-R* framework, which has several models of serial recall (e.g., Anderson, 1972) relevant to the number of recalled notes and emphasises modelling produced events, we intend to more thoroughly explore modelling that integrates the *ACT-R* framework (Ritter, Tehranchi, & Oury, 2019) alongside melodic feature modelling. We note that integrations of musicological considerations with *ACT-R* seem to be scarce (Chikhaoui et al., 2009; Reiter-Haas et al., 2021), yet such a production-driven modelling framework seems highly relevant to musical phenomena (Okada & Slevc, 2021).

# References

Anderson, J. R. (1972). Fran: A simulation model of free recall. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 5, pp. 315–378). Academic Press. <https://doi.org/10.1016/S0079-7421(08)60444-2>

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. <https://doi.org/10.1016/S1364-6613(00)01538-2>

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press. <https://doi.org/10.1016/S0079-7421(08)60452-1>

Baker, D. (2019). Modeling melodic dictation. *LSU Doctoral Dissertations*. <https://doi.org/10.31390/gradschool_dissertations.4960>

Baker, D. (2021). MeloSol corpus. *Empirical Musicology Review*, *16*, 106–113. <https://doi.org/10.18061/emr.v16i1.7645>

Berz, W. L. (1995). Working memory in music: A theoretical model. *Music Perception*, *12*(3), 353–364. <https://doi.org/10.2307/40286188>

Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130. <https://doi.org/10.1016/j.cognition.2005.11.007>

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion*, *19*(8), 1113–1139. <https://doi.org/10.1080/02699930500204250>

Buren, V., Müllensiefen, D., Roeske, T., & Degé, F. (2021). What makes a child musical? Conceptions of musical ability in childhood. *Early Child Development and Care*, *191*(12), 1985–2000. <https://doi.org/10.1080/03004430.2020.1866566>

Chenette, T. (2021). What are the truly aural skills? *Music Theory Online*, *27*(2). Retrieved from <https://mtosmt.org/issues/mto.21.27.2/mto.21.27.2.chenette.html>

Chikhaoui, B., Ene, H., Beaudoin, M., Pratte, G., Bellefeuille, P., & Laudares, F. (2009). Learning a song: An ACT-r model. *World Academy of Science, Engineering and Technology 31 Proceedings*. Presented at the World academy of science, engineering and technology.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*. <https://doi.org/10.1017/S0140525X1500031X>

Clayton, M. (2008). The social and personal functions of music in cross-cultural perspective. In *Oxford handbook of music psychology*. <https://doi.org/10.1093/oxfordhb/9780199298457.013.0004>

Clemente, A., Vila-Vidal, M., Pearce, M. T., Aguiló, G., Corradi, G., & Nadal, M. (2020). A set of 200 musical stimuli varying in balance, contour, symmetry, and complexity: Behavioral and computational assessments. *Behavior Research Methods*, *52*(4), 1491–1509. <https://doi.org/10.3758/s13428-019-01329-8>

Cornelius, N., & Brown, J. L. (2020). The interaction of repetition and difficulty for working memory in melodic dictation tasks. *Research Studies in Music Education*, *42*(3), 368–382. <https://doi.org/10.1177/1321103X18821194>

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. <https://doi.org/10.1177/0963721409359277>

Cox, A. (2001). The mimetic hypothesis and embodied musical meaning. *Musicae Scientiae*, *5*(2), 195–212. <https://doi.org/10.1177/102986490100500204>

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics*, *28*(5), 381–389. <https://doi.org/10.3758/BF03204881>

Dewitt, L. A., & Crowder, R. G. (1986). Recognition of novel melodies after brief delays. *Music Perception: An Interdisciplinary Journal*, *3*(3), 259–274. <https://doi.org/10.2307/40285336>

Dowling, W. J. (1972). Recognition of melodic transformations: Inversion, retrograde, and retrograde inversion. *Perception & Psychophysics*, *12*(5), 417–421. <https://doi.org/10.3758/BF03205852>

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*(4), 341–354. <https://doi.org/10.1037/0033-295X.85.4.341>

Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception & Psychophysics*, *50*(4), 305–313. <https://doi.org/10.3758/bf03212222>

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

Dowling, W. J., Kwak, S., & Andrews, M. W. (1995). The time course of recognition of novel melodies. *Perception & Psychophysics*, *57*(2), 136–149. <https://doi.org/10.3758/bf03206500>

Dowling, W., & Bartlett, J. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology: A Journal of Research in Music Cognition*, *1*. <https://doi.org/10.1037/h0094275>

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, *37*(1), 295–340. <https://doi.org/10.1002/aris.1440370108>

Dreyfus, L., Crawford, T., Müllensiefen, D., & Baker, D. (2016). Recognition of leitmotives in richard wagner’s music: An item response theory approach. In A. F. X. Wilhelm & H. A. Kestler (Eds.), *Analysis of large and complex data* (pp. 473–483). Springer International Publishing. Retrieved from <https://www.springer.com/gb/book/9783319252247>

Edworthy, J. (1985). Interval and contour in melody processing. *Music Perception: An Interdisciplinary Journal*, *2*(3), 375–388. <https://doi.org/10.2307/40285305>

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. <https://doi.org/10.1037/0033-295x.102.2.211>

Ettlinger, M., Margulis, E., & Wong, P. (2011). Implicit memory in music and language. *Frontiers in Psychology*, *2*. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2011.00211>

Frieler, K. (2018). Computational melody analysis. In M. Pfleiderer, K. Frieler, W.-G. Abeßer, B. Zaddach, & B. Burkhard (Eds.), *Inside the jazzomat. New perspectives for jazz research* (pp. 41–84). Schott Campus.

Gates, S. (2021). Developing musical imagery: Contributions from pedagogy and cognitive science. *Music Theory Online*, *27*(2). Retrieved from <https://mtosmt.org/issues/mto.21.27.2/mto.21.27.2.gates.html>

Gobet, F. (2005). Chunking models of expertise: Implications for education. *Applied Cognitive Psychology*, *19*(2), 183–204. <https://doi.org/10.1002/acp.1110>

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. <https://doi.org/10.1016/S1364-6613(00)01662-4>

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511574931>

Hallam, S., & Creech, A. E. (2010). *Music education in the 21st century in the united kingdom: Achievements, analysis and aspirations*. Institute of Education - London.

Harrison, P. M. C., Bianco, R., Chait, M., & Pearce, M. T. (2020). PPM-decay: A computational model of auditory prediction with memory decay. *PLOS Computational Biology*, *16*(11), e1008304. <https://doi.org/10.1371/journal.pcbi.1008304>

Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *7*(1), 3618. <https://doi.org/10.1038/s41598-017-03586-z>

Harrison, P. M. C., Musil, J. J., & Müllensiefen, D. (2016). Modelling melodic discrimination tests: Descriptive and explanatory approaches. *Journal of New Music Research*, *45*(3), 265–280. <https://doi.org/10.1080/09298215.2016.1197953>

Herborn, P. (2022). *Features of the perception and construction of melodies*. Emanobooks.

Idson, W. L., & Massaro, D. W. (1978). A bidimensional model of pitch in the recognition of melodies. *Perception & Psychophysics*, *24*(6), 551–565. <https://doi.org/10.3758/BF03198783>

Karpinski, G. S. (2000). *Aural skills acquisition: The development of listening, reading, and performing skills in college-level musicians*. Oxford University Press.

Kauffman, W. H., & Carlsen, J. C. (1989). Memory for intact music works: The importance of music expertise and retention interval. *Psychomusicology: A Journal of Research in Music Cognition*, *8*(1), 3–20. <https://doi.org/10.1037/h0094235>

Koh, C. (2002). *Memory and learning in music reproduction: The effects of melodic structure, perceptual cues and learning methods on music recall* (PhD thesis).

Krumhansl, C. (1990). *Cognitive foundations of musical pitch*. Oxford University Press, USA.

Lehmann, A. C., Sloboda, J. A., & Woody, R. H. (2007). *Psychology for musicians: Understanding and acquiring the skills*. Oxford University Press, USA.

Long, J. D. (2011). *Longitudinal data analysis for the behavioral sciences using r*. SAGE Publications.

Long, P. A. (1977). Relationships between pitch memory in short melodies and selected factors. *Journal of Research in Music Education*, *25*(4), 272–282. <https://doi.org/10.2307/3345268>

Massaro, D., Kallman, H., & Kelly, J. (1980). The role of tone height, melodic contour, and tone chroma in melody recognition. *Journal of Experimental Psychology. Human Learning and Memory*, *6*, 77–90. <https://doi.org/10.1037//0278-7393.6.1.77>

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. <https://doi.org/10.1037/h0043158>

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*(3), 161–175. Retrieved from <https://www.jstor.org/stable/30200223>

Müllensiefen, D. (2009). *FANTASTIC: Feature ANalysis technology accessing STatistics (in a corpus; technical report)*. 37. Retrieved from <http://www.doc.gold.ac.uk/isms/m4s/FANTASTIC_docs.pdf>

Müllensiefen, D., & Frieler, K. (2004a). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. Human judgments. In W. B. Hewlett & E. Selfridge-Field, *Music query: Methods, models, and user studies*. MIT Press.

Müllensiefen, D., & Frieler, K. (2004b). Melodic similarity: Approaches and applications. In S. D. Lipscombe, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the 8th ICMPC* (pp. 283–289). IL.

Müllensiefen, D., & Frieler, K. (2007). Modelling experts’ notions of melodic similarity. *Musicae Scientiae*, *11*(1), 183–210. <https://doi.org/10.1177/102986490701100108>

Müllensiefen, D., & Halpern, A. R. (2014). The role of features and context in recognition of novel melodies. *Music Perception*, *31*(5), 418–435. <https://doi.org/10.1525/mp.2014.31.5.418>

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

Müllensiefen, D., & Wiggins, G. A. (2011). Sloboda and parker’s recall paradigm for melodic memory: A new, computational perspective. In *Music and the mind: Essays in honour of john sloboda* (pp. 161–186). New York, NY, US: Oxford University Press.

Murdock Jr., B. B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology*, *60*(4), 222–234. <https://doi.org/10.1037/h0045145>

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. <https://doi.org/10.1111/j.2041-210x.2012.00261.x>

Oberauer, K., & Cowan, N. (2007). Working memory capacity. *Experimental Psychology*, *54*, 245–246. <https://doi.org/10.1027/1618-3169.54.3.245>

Ogawa, Y., Kimura, T., & Mito, H. (1995). Modification of musical schema for japanese melody: A study of comprehensible and memorable melody. *Bulletin of the Council for Research in Music Education*, (127), 136–141. Retrieved from <https://www.jstor.org/stable/40318777>

Okada, B. M., & Slevc, R. (2021). What is “musical ability” and how do we measure it? *Proceedings of the Future Directions of Music Cognition International Conference*. Presented at the Music cognition international conference.

Oura, Y., & Hatano, G. (1988). Memory for melodies among subjects differing in age and experience in music. *Psychology of Music*, *16*(2), 91–109. <https://doi.org/10.1177/0305735688162001>

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395. <https://doi.org/10.1111/nyas.13654>

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). Melodic grouping in music information retrieval: New methods and applications. In Z. W. Raś & A. A. Wieczorkowska (Eds.), *Advances in music information retrieval* (pp. 364–388). Berlin, Heidelberg: Springer. <https://doi.org/10.1007/978-3-642-11674-2_16>

Pearce, M., & Müllensiefen, D. (2017). Compression-based modelling of musical similarity perception. *Journal of New Music Research*, *46*(2), 135–155. <https://doi.org/10.1080/09298215.2017.1305419>

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.-G., & Burkhart, B. (Eds.). (2017). *Inside the jazzomat - new perspectives for jazz research*. Schott Campus.

Reiter-Haas, M., Parada-Cabaleiro, E., Schedl, M., Motamedi, E., Tkalcic, M., & Lex, E. (2021). Predicting music relistening behavior using the ACT-r framework. In *Fifteenth ACM conference on recommender systems* (pp. 702–707). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3460231.3478846>

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-r: A cognitive architecture for modeling cognition. *WIREs Cognitive Science*, *10*(3), e1488. <https://doi.org/10.1002/wcs.1488>

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from <https://www.jstatsoft.org/v48/i02/>

Salakka, I., Pitkäniemi, A., Pentikäinen, E., Mikkonen, K., Saari, P., Toiviainen, P., & Särkämö, T. (2021). What makes music memorable? Relationships between acoustic musical features and music-evoked emotions and memories in older adults. *PLOS ONE*, *16*(5), e0251692. <https://doi.org/10.1371/journal.pone.0251692>

Schellenberg, E. G., Weiss, M. W., Peng, C., & Alam, S. (2019). Fine-grained implicit memory for key and tempo. *Music & Science*, *2*, 2059204319857198. <https://doi.org/10.1177/2059204319857198>

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, *101*(2), 357–361. <https://doi.org/10.1037/0033-295X.101.2.357>

Silas, S., Müllensiefen, D., Gelding, R., Frieler, K., & Harrison, P. M. C. (2022). The associations between music training, musical working memory, and visuospatial working memory: An opportunity for causal modeling. *Music Perception*, *39*(4), 401–420. <https://doi.org/10.1525/mp.2022.39.4.401>

Silas, S., Müllensiefen, D., & Kopiez, R. (2023). Singing ability assessment: Development and validation of an open-source test environment for singing data. *Behaviour Research Methods. In Review.*

Silas, S., Robinson, M., Baker, D., Müllensiefen, D., Harrison, P., & Jacoby, N. (2023). The memorability of sonic logos: A multi-paradigm approach. *In Prep.*

Sloboda, J. A. (1985). *The musical mind: The cognitive psychology of music*. Clarendon Press.

Sloboda, J. A., & Parker, D. H. H. (1985). Immediate recall of melodies. In R. West, P. Howell, & I. Cross (Eds.), *Musical structure and cognition* (pp. 143–167). London: Academic Press.

Sturm, B. L. (2013). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, *41*(3), 371–406. <https://doi.org/10.1007/s10844-013-0250-y>

Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PLOS ONE*, *12*(10), e0186773. <https://doi.org/10.1371/journal.pone.0186773>

Talamini, F., Carretti, B., & Grassi, M. (2016). The working memory of musicians and nonmusicians. *Music Perception: An Interdisciplinary Journal*, *34*(2), 183–191. <https://doi.org/10.1525/mp.2016.34.2.183>

Tarr, B., Launay, J., & Dunbar, R. I. M. (2014). Music and social bonding: “Self-other” merging and neurohormonal mechanisms. *Frontiers in Psychology*, *5*. <https://doi.org/10.3389/fpsyg.2014.01096>

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 37–55. <https://doi.org/10.1037/xlm0000578>

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*(4), 885–913. <https://doi.org/10.1037/0033-295X.107.4.885>

Typke, R., Wiering, F., & Veltkamp, R. C. (2007). Transportation distances and human perception of melodic similarity. *Musicae Scientiae*, *11*(1), 153–181. <https://doi.org/10.1177/102986490701100107>

Uitdenbogerd, A. (2002). *Music information retrieval technology* (PhD thesis). Retrieved from <http://www.pampalk.at/mir-phds/abstract/Uitdenbogerd2002.html>

Vergauwe, E., Barrouillet, P., & Camos, V. (2010). Do mental processes share a domain-general resource? *Psychological Science*, *21*(3), 384–390. <https://doi.org/10.1177/0956797610361340>

Yuan, Y., Oishi, S., Cronin, C., Müllensiefen, D., Atkinson, Q., Fujii, S., & Savage, P. E. (2020). *Perceptual vs. Automated judgments of music copyright infringement*. PsyArXiv. <https://doi.org/10.31234/osf.io/tq7v5>

Zielinska, H., & Miklaszewski, K. (1992). Memorising two melodies of different style. *Psychology of Music*, *20*(2), 95–111. <https://doi.org/10.1177/0305735692202001>

# Appendices

# Appendix A

## Dependent variables: Our main similarity variables of interest are highlighted in orange.

| Measure | Definition |
| --- | --- |
| ngrukkon | Ukkonen measures for n-grams on raw pitch values |
| harmcore | Edit Distance of harmonic symbols per segment, obtained via Krumhansl’s tonality vectors. |
| rhythfuzz | Edit distance of classified length of melody tones. |
| opti3 | 3.027 \* ngrukkon + 2.502 \* rhythfuzz + 1.439 \* harmcore - 0.146 |
| No. Recalled Notes | The number of notes the participant produced in the trial. |
| No. Correct | The number of correct notes the participant sang (allowing octave errors). |
| No. Errors | The number of errors participants sang (allowing octave errors). |

# Appendix B

## Examples showing development of trial performance and a qualitative description of their change in similarity.

To illustrate the similarity measurement for different musical dimensions, the following figures show a typical improvement of the sung recalls of one participant of a particular test item. Due to space limitations, only the transcriptions of two trials can be depicted here.

Participant *F.S.* (male, aged 29, plays bass in rock band) listened to an audio excerpt from the pop song “Cold, Cold Heart” by *Wet Wet Wet*, the vocal melody of which is depicted in Figure 9. Despite the fact that it contains a large jump of one and a half octaves, participants found this melody generally rather easy to reproduce from memory, probably due to its clear phrase structure. Figure 10 shows the recall of *F.S*. after the third listening. The rendition resembles most to the third (and last) phrase of the original melody. Figure 11 depicts the rendition on the sixth and last trial. A great improvement in accuracy is clearly visible, although he altered all B flats to B naturals. When reflected by similarity algorithms, this will generally show an increasing score in overall similarity per trial, though perhaps differing by specific dimension, depending on the type of error made in a particular attempt.



Figure3: Transcription of original melody from chorus of “Cold Cold Heart” by Wet Wet Wet



Figure4: Transcription of F.S.’s sung recall of the test item on the third trial



Figure5: Transcription of F.S.’s sung recall of the test item on the sixth trial

# Appendix C

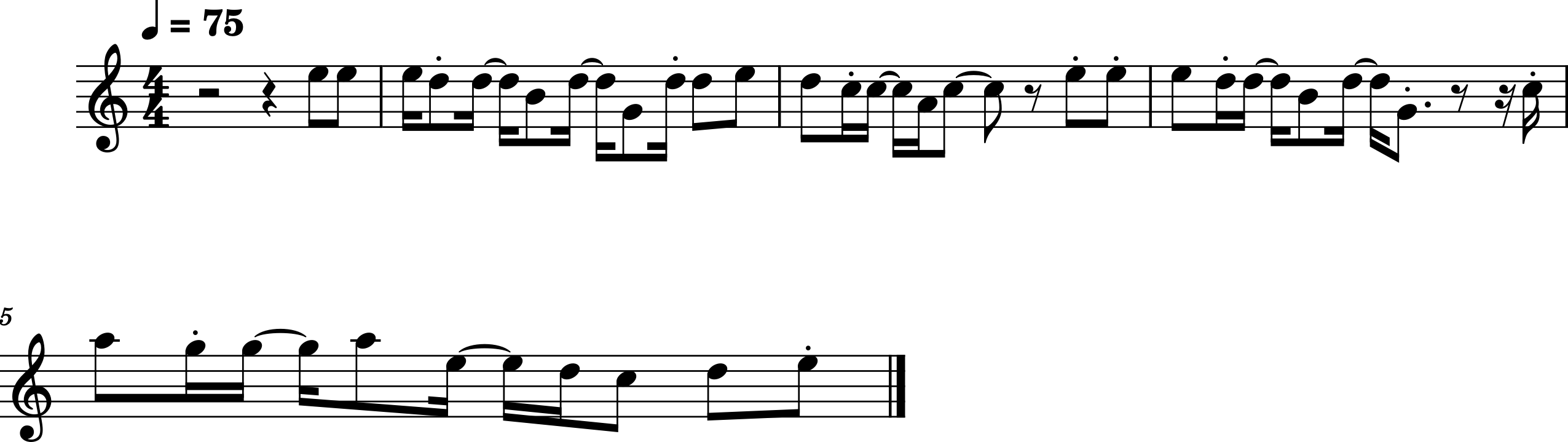
## Short melodic excerpts from pop songs used as materials in the study

| No. | Song | Composer/Interpreter | Genre/Meter | Tempo |
| --- | --- | --- | --- | --- |
| 1 | Children Of The Night | R. Marx | Pop 4/4 | 75 |
| 2 | Climb Up | N. Sedaka | R´n´R 4/4 | 120 |
| 3 | Cold Cold Heart | M. Pellow | Pop 4/4 | 80 |
| 4 | Do You Want To Dance? | R. Freeman | R´n´R 4/4 | 100 |
| 5 | Du gehörst zu mir | J. Heider | Schlager 4/4 | 120 |
| 6 | Longer | D. Fogelberg | Pop-Ballade 4/4 | 80 |
| 7 | Oh Carol | N. Sedaka | Pop-Ballade 4/4 | 140 |
| 8 | Take Good Care | C. King | Ballade 4/4 | 120 |
| 9 | The Sky Is Crying | M. Levy | Blues 12/8 | 60 |
| 10 | You Are My Destiny | P. Anka | Schlager 12/8 | 85 |
| 11 | Goodbye My Love Goodbye | M. Panas / D. Roussos | Schlager 4/4 | 114 |
| 12 | Enjoy Your Life | Funky Be | HipHop 4/4 | 85 |
| 13 | Love Is Like A Rainbow | T. Anders | Disco-Pop 4/4 | 155 |
| 14 | Let Me Be Your Only One | Funky Be | HipHop 4/4 | 100 |

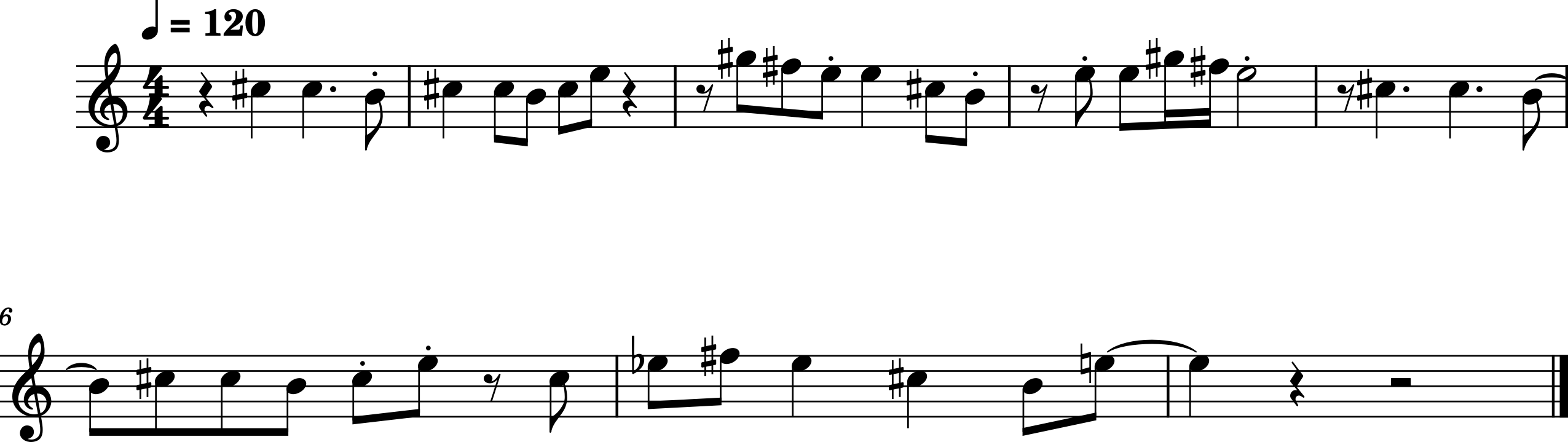
## Appendix C1

### An example of one melody from each pop song..

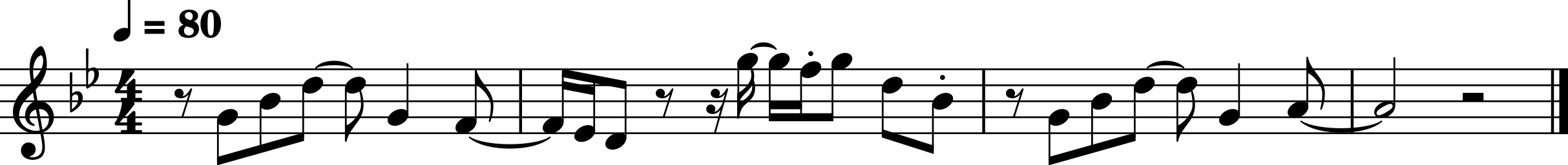
#### Melody No. 1: Children of the Night, R.Marx.



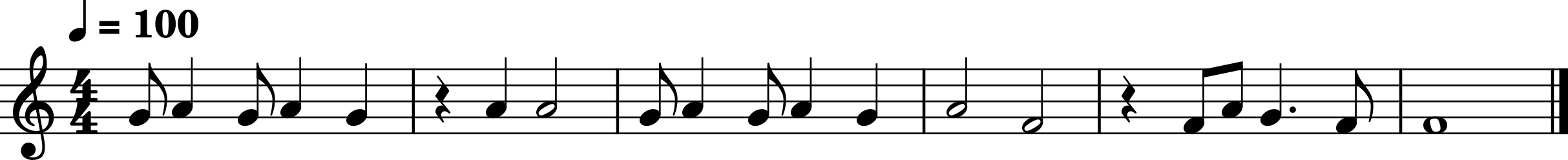
#### Melody No. 2: Climb Up, N.Sedaka.



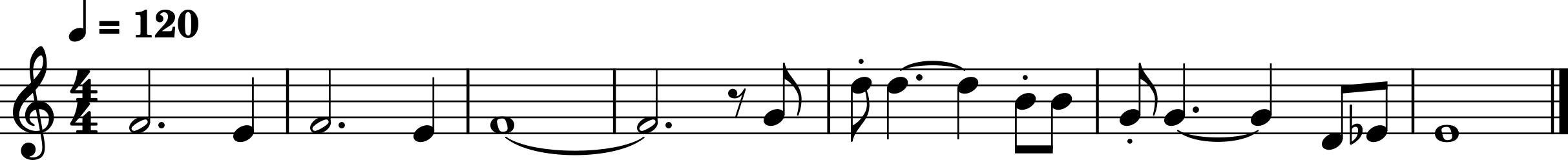
#### Melody No. 3: Cold Cold Heart, M.Pellow.



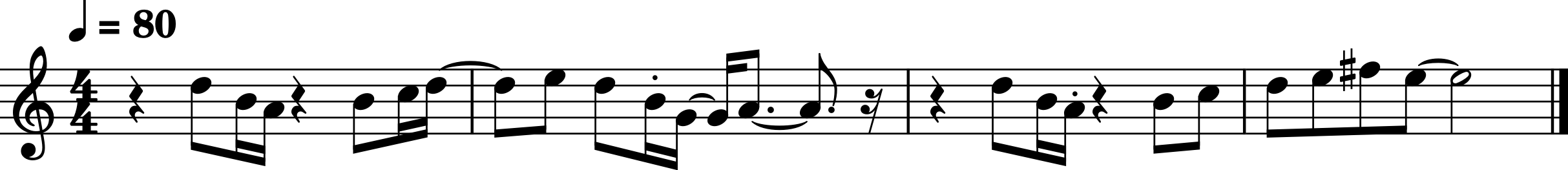
#### Melody No. 4: Do You Want To Dance?, R. Freeman.



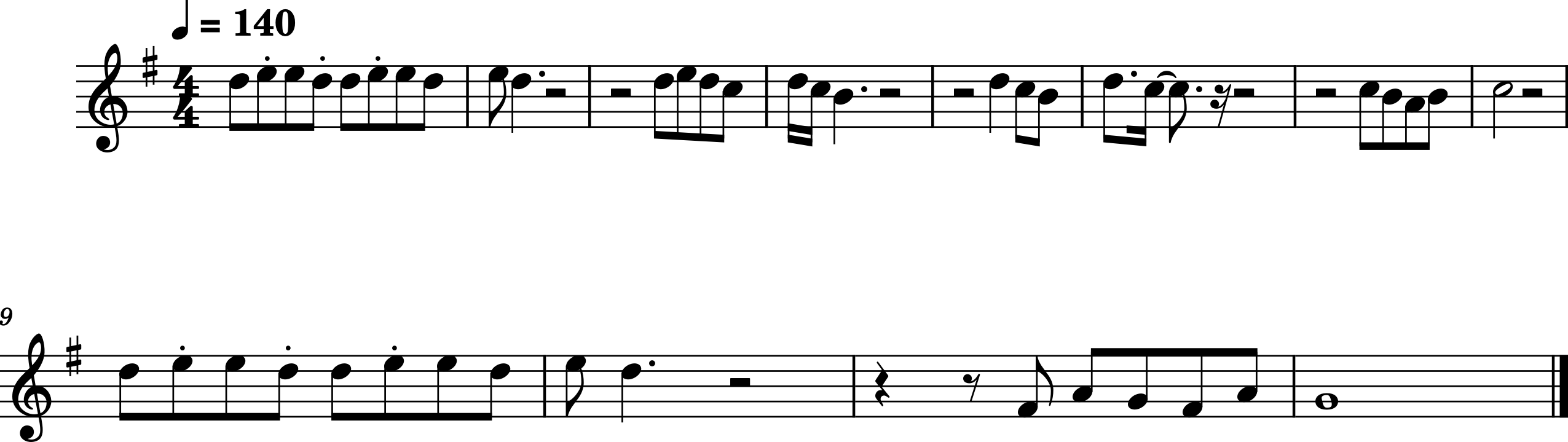
#### Melody No. 5: Du gehörst zu mir, J. Heider.



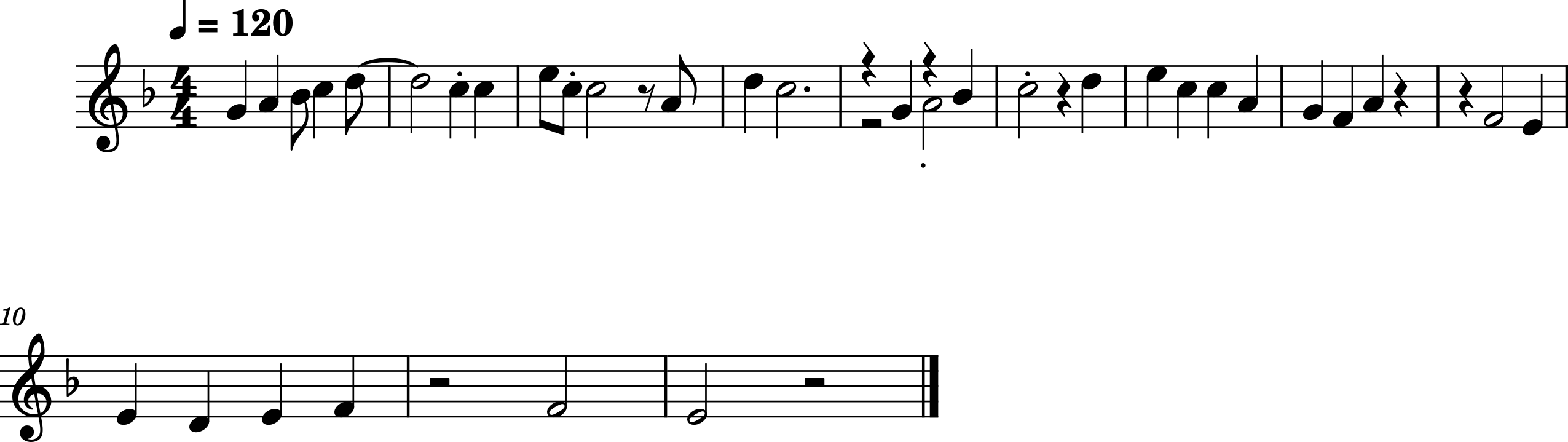
#### Melody No. 6: Longer, D.Fogelberg.



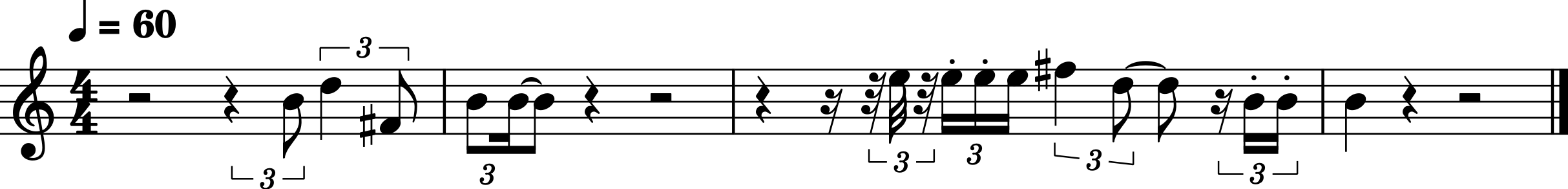
#### Melody No. 7: Oh Carol, N. Sedaka.



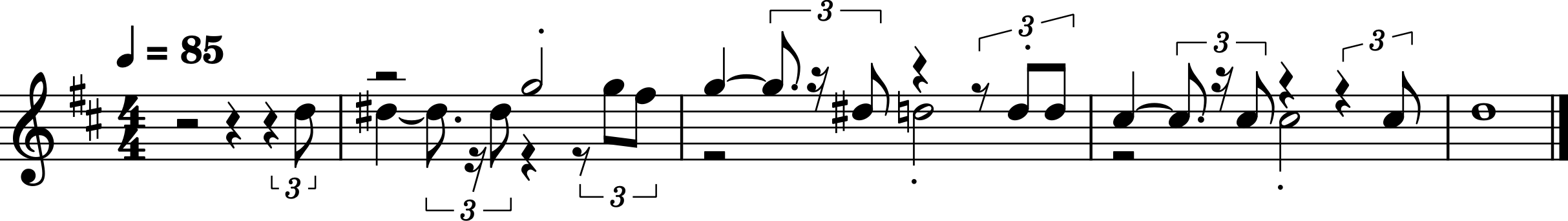
#### Melody No. 8: Take Good Care, C. King.



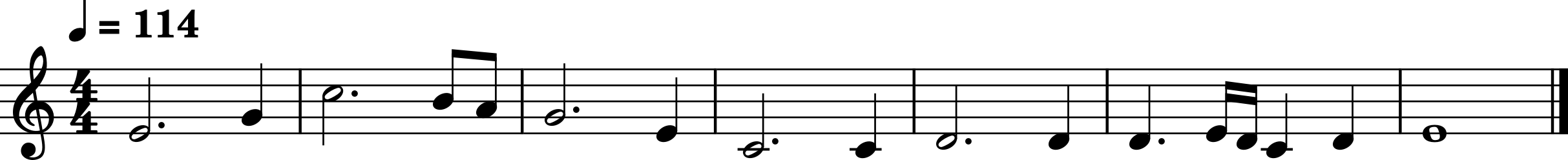
#### Melody No. 9: The Sky is Crying, M.Levy.



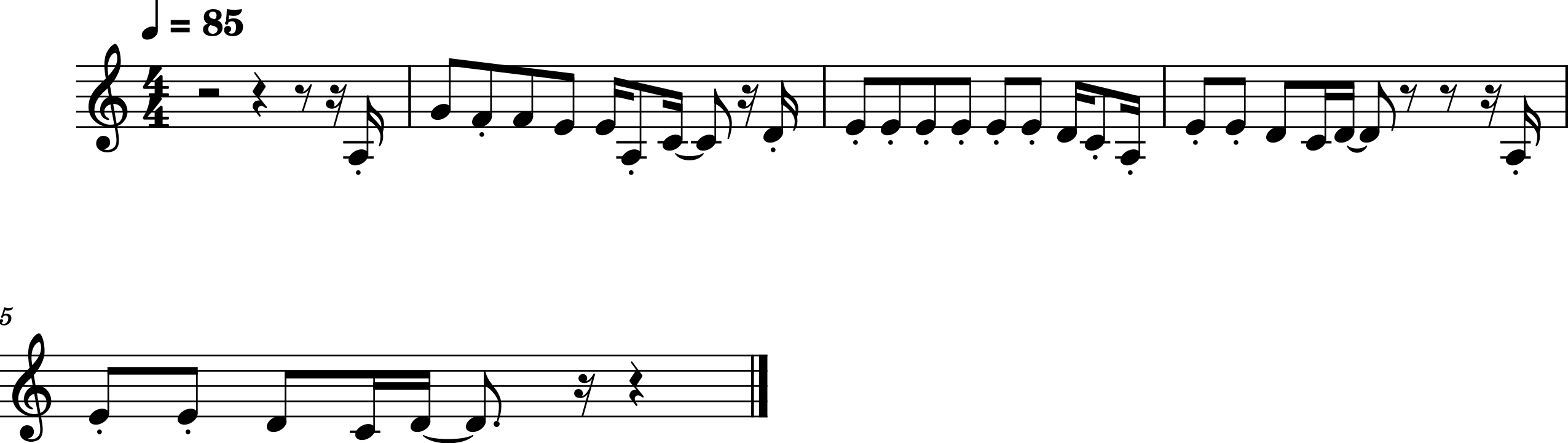
#### Melody No. 10: You Are My Destiny, P. Anka.



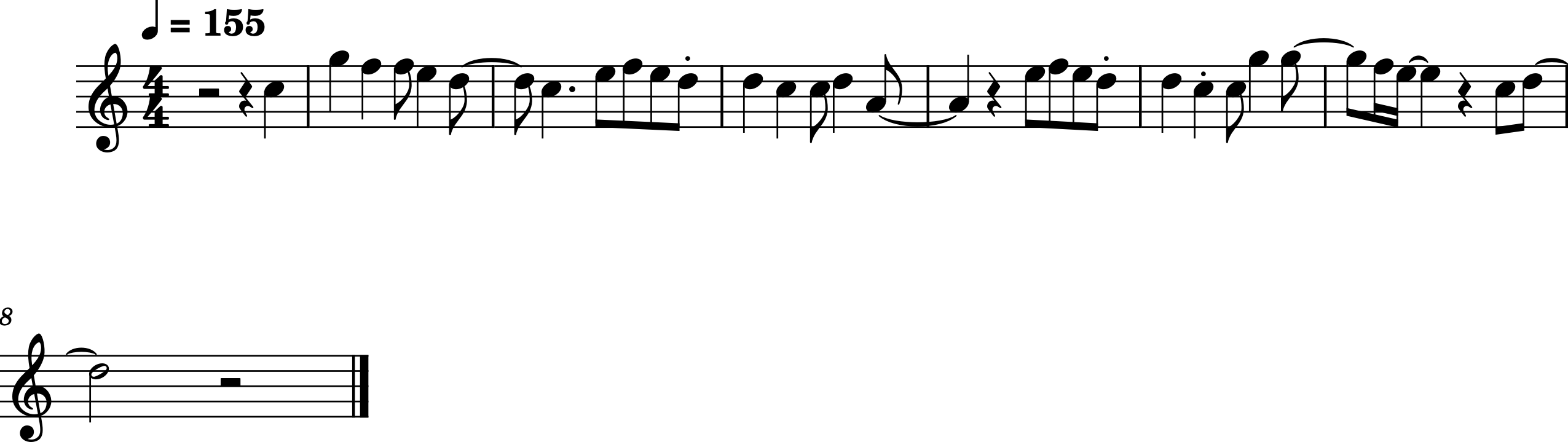
#### Melody No. 11: Goodbye My Love Goodbye, M. Panas / D. Roussos.



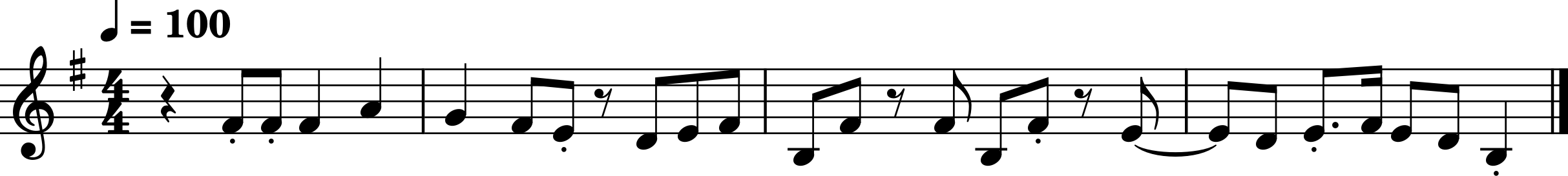
#### Melody No. 12 Enjoy Your Life, Funky Be.



#### Melody No. 13: Love Is Like A Rainbow, T. Anders.



#### Melody No. 14: Let Me Be Your Only One, Funky Be.



## Appendix C2

### Description and distribution of melodic features.

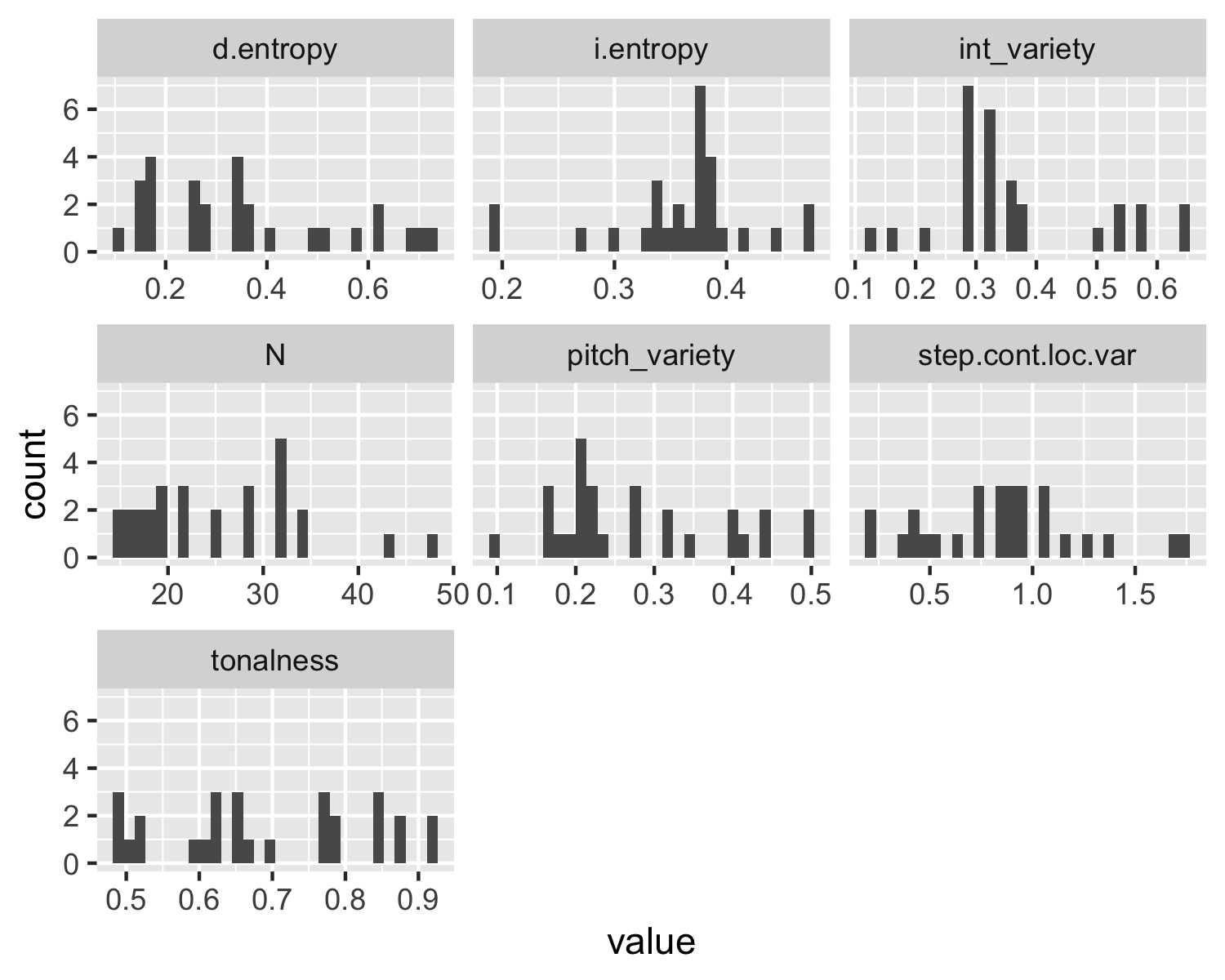


Table 7: Melodic feature summary statistics. Note, some are not used in our modelling, but are here to show other properties of the melodies.

| Feature | Mean | SD | Coefficient of Variation |
| --- | --- | --- | --- |
| d.entropy | 0.36 | 0.19 | 0.53 |
| step.cont.loc.var | 0.86 | 0.38 | 0.44 |
| pitch.variety | 0.28 | 0.11 | 0.40 |
| mean.int.size | 2.21 | 0.84 | 0.38 |
| int.variety | 0.37 | 0.13 | 0.37 |
| N | 25.39 | 8.66 | 0.34 |
| mean.information.content | 4.07 | 0.89 | 0.22 |
| tonalness | 0.69 | 0.14 | 0.20 |
| i.entropy | 0.36 | 0.06 | 0.18 |

# Appendix D1

## Questionnaire items.

| Variable | Question | Response Format |
| --- | --- | --- |
| chorusin | Do you sing in a choir? | Yes/No |
| singinstr: | Have you ever received singing instructions? | Yes/No |
| yearsins | For how many years have you been playing an instrument or making music? | \_\_years |
| musmakpa | During your most active musical phase how many hours per week did you make music (practice+rehearsal+gigs+lessons+playing+etc.) | \_\_hours/week |
| paidless | For how many months have you received paid instrumental or singing lessons? | \_\_ months |
| paidgigs | How many gigs have you played that you have been paid for? | \_\_\_gigs |
| gigs | Overall, how many gigs have you played in front of an audience in your life? | \_\_\_gigs |

# Appendix D2

## Factor loadings for mixed type variables based on questionnaire items

| Variable | Loading | h2 | u2 |
| --- | --- | --- | --- |
| chorusin | 0.70 | 0.49 | 0.51 |
| singinstr | 0.46 | 0.21 | 0.79 |
| yearsins | 0.82 | 0.67 | 0.33 |
| musmakpa | 0.77 | 0.59 | 0.41 |
| paidless | 0.65 | 0.42 | 0.58 |
| paidgigs | 0.66 | 0.44 | 0.56 |
| gigs | 0.69 | 0.48 | 0.52 |

# Appendix E

## Melodic features used in this paper

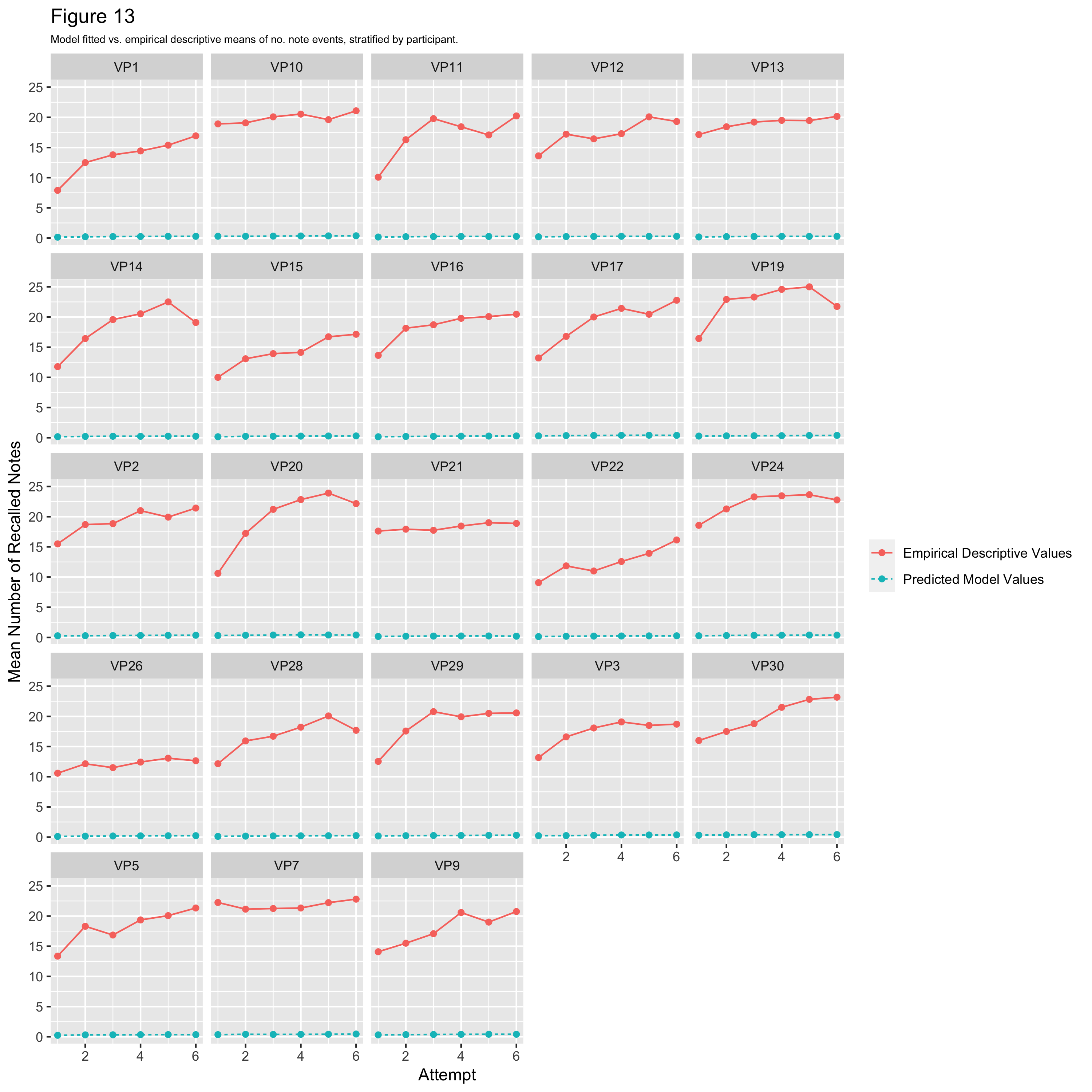
| Feature | Description | Equation | Reference |
| --- | --- | --- | --- |
| N | The length of the target melody. | - | - |
| log\_freq | The log of the relative count of a frequency in the corpus | - | - |
| i.entropy | The average level of “information” or “surprise” in intervallic representations. Specifically, a variant of Shannon entropy on interval representations (Shannon, 1948) |  | Müllensiefen, 2009 |
| step.cont.loc.var | The mean absolute difference between adjacent values in the vector representing of step contour. |  | Müllensiefen, 2009 |
| d.entropy | The average level of “information” or “surprise” in rhythm values. Specifically, a variant of Shannon entropy on rhythmic representations (Shannon, 1948) |  | Müllensiefen, 2009 |
| mean\_information\_content | The average information content contained in the pitch values of a melody. Can be thought of as quantifying a melody’s self-similarity. | - | Harrison, Bianco, Chait & Pearce (2020) |

# Appendix F1

## Average by-participant across trials

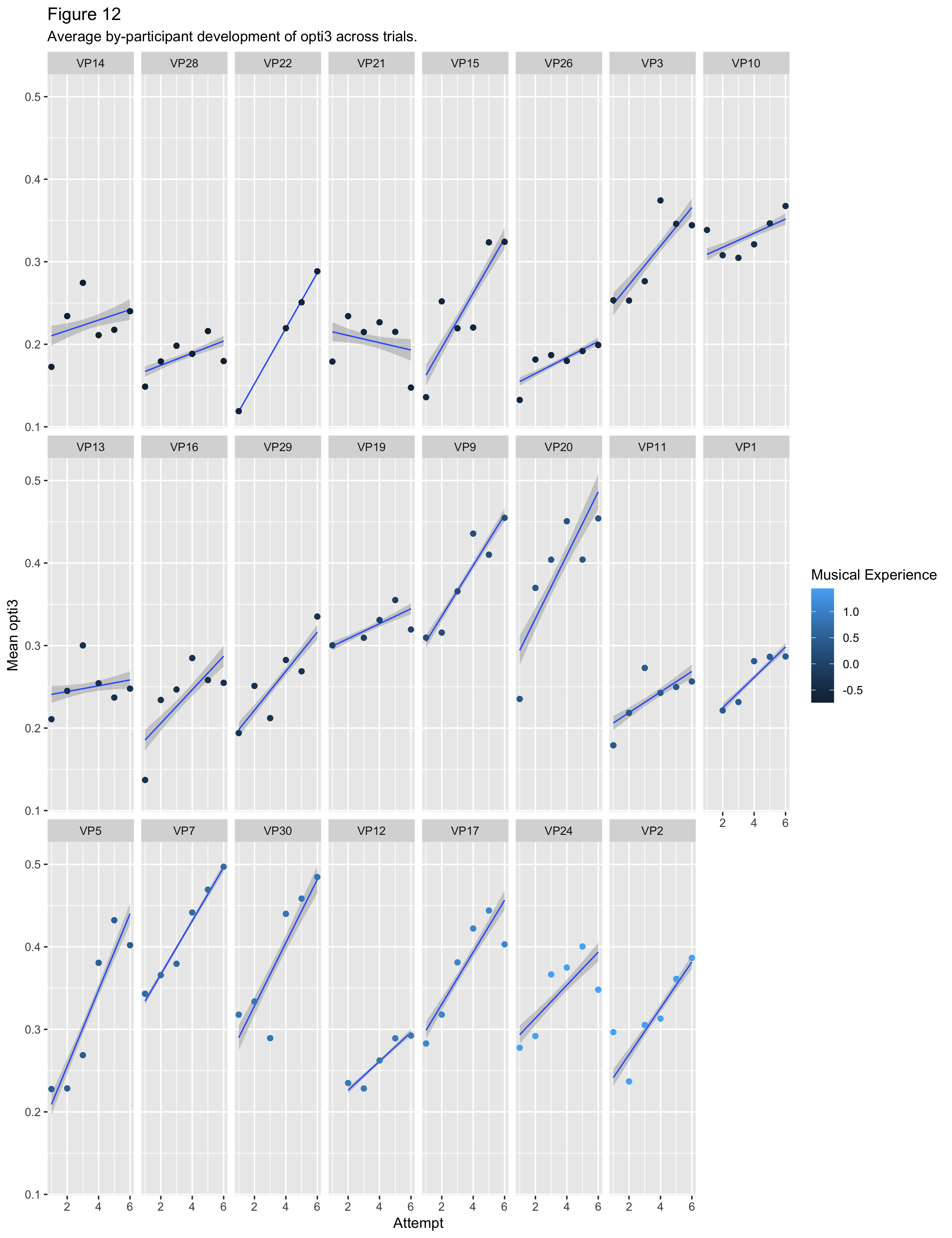
Another way of visualising differences in performance is at the level of participant, coloured and ordered by level of musical experience. This is useful since it invokes no false dichotimisations and preserves the actual unit of participant (however, note that participant-level effects are captured by our mixed effects models).

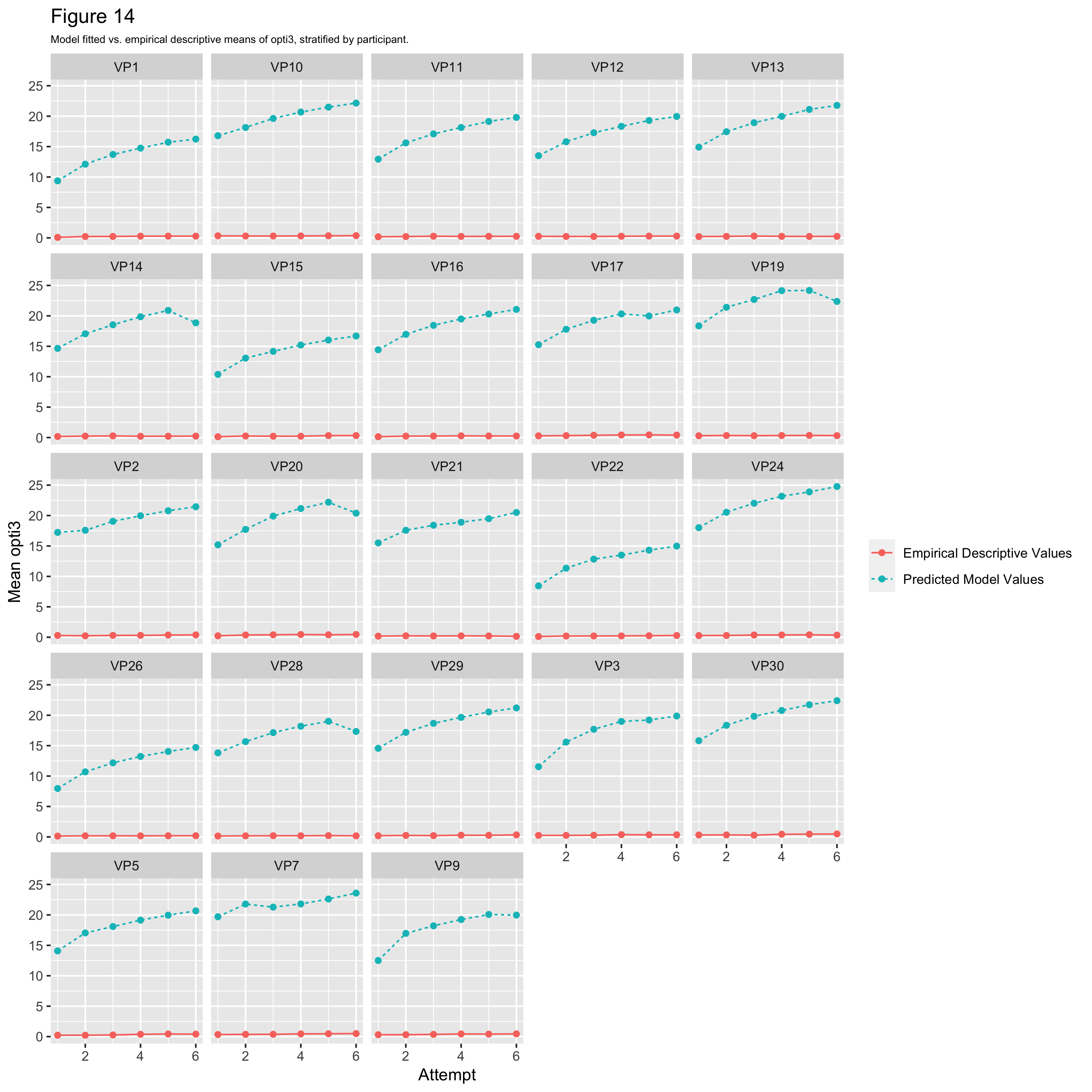
## Average by-participant development of number of recalled notes events across trials



## Average by-participant development of opti3 across trials

As shown in the figure, participants seem to have vastly different slopes. The bottom right, lighter blue, higher musical experience participants (e.g., *VP5*, *VP7*, *VP30*, *VP12*, *VP17*, *VP24*, *VP2*) seem to have steeper slopes than the lower musical experience participants in the top left, darker coloured (*VP14*, *VP28*, *VP21*), suggesting that higher musical experience is related to quicker learning. However, note that this pattern is not the same for everyone e.g., *VP22* has a steep slope, but scores low on musical experience



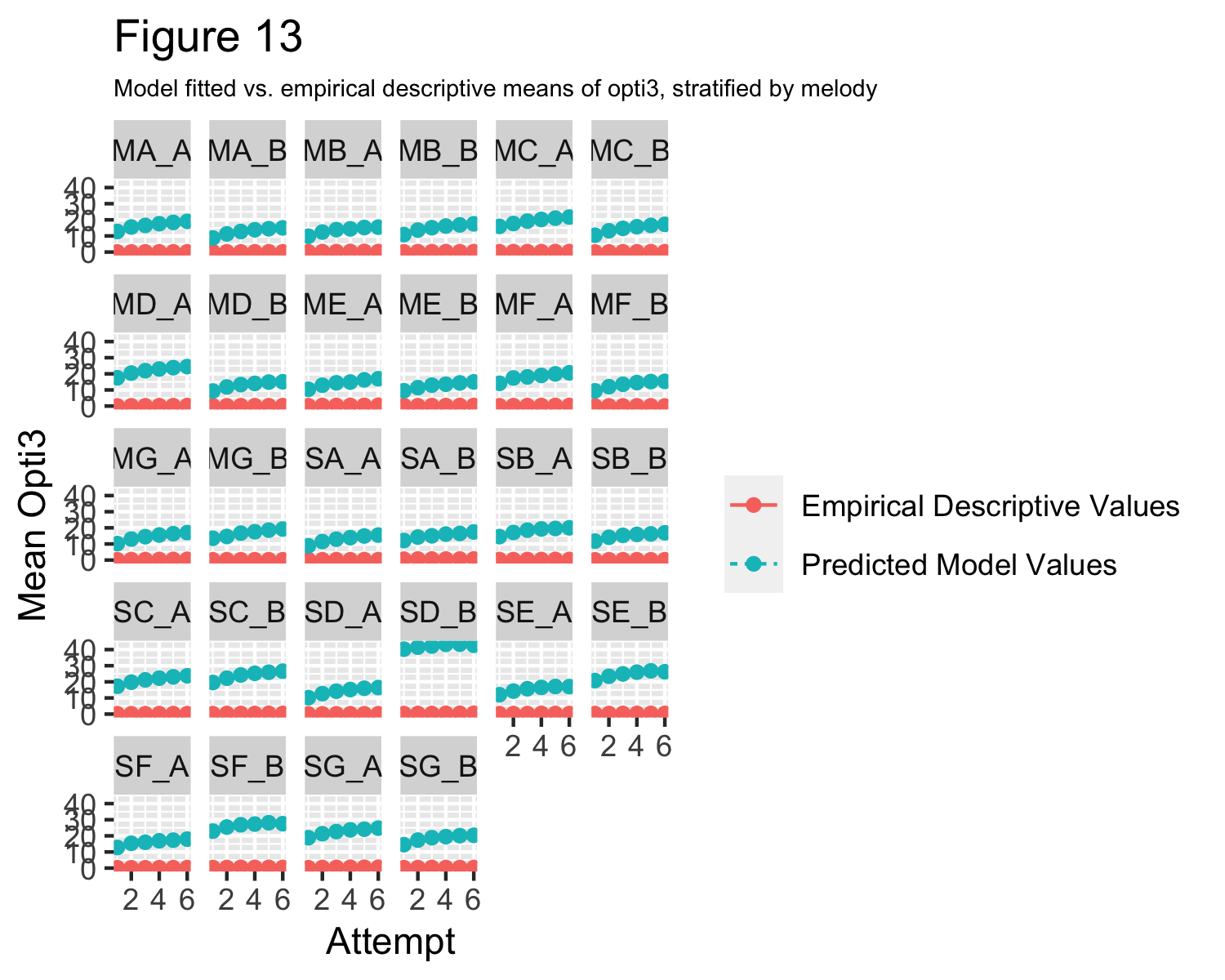


# Appendix F2

## Average by-melody development of no. of note events across trials

Melody SA\_B appears to be the easiest melody to recall (mean *opti3* across all trials = 0.61), whereas MF\_B appears most difficult to recall (mean *opti3* across all trials = 0.11). This shows that there can be substantial variation in the difficulty of each melody.





# Appendix G Linear vs. Non-Linear Models for number of recalled notes and opti3

We proceed by using the log attempt as numeric predictor, owing to the observed non-linearities in both *opti3* and *number of recalled notes* across attempt. A comparison of linear vs non-linear models is shown below.

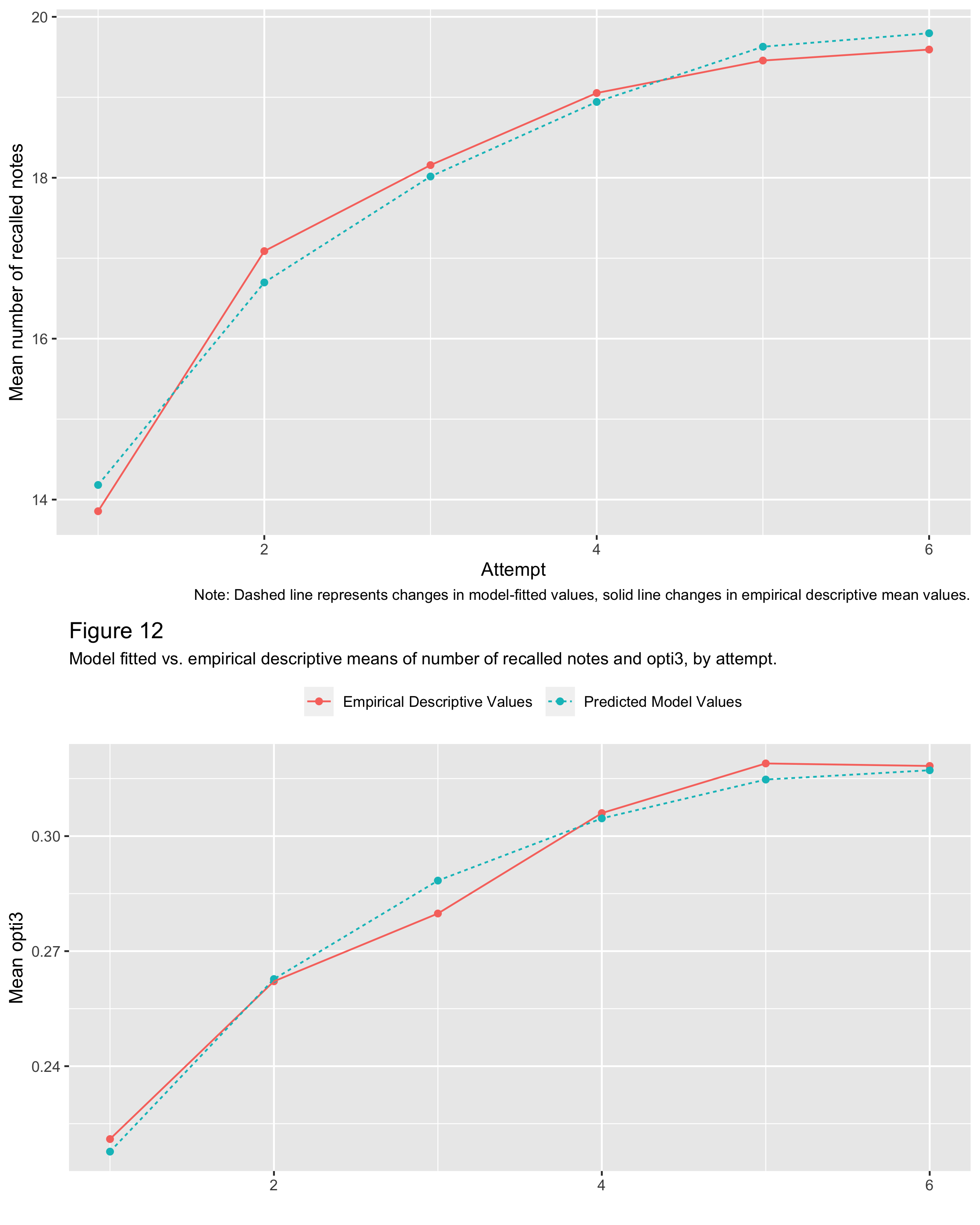
## Appendix G1: Linear model of mean similarity scores (number of recalled notes) across repeated attempts

Table 8:

\*\*

| Term |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 13.89 | [11.30, 16.49] | 10.49 | 37.93 | < .001 |
| Attempt | 1.22 | [1.10, 1.33] | 20.14 | 1,466.66 | < .001 |

Figure 12 shows that the use of the log attempt as predictor is justified, capturing the systematic non-linear pattern generally well.



## Appendix G2: Linear model of mean similarity scores (opti3) across repeated attempts

Table 9:

\*\*

| Term |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 0.21 | [0.15, 0.26] | 7.57 | 44.26 | < .001 |
| Attempt | 0.03 | [0.02, 0.03] | 15.55 | 1,460.28 | < .001 |

# Appendix H

## Diagnostic statistics for models with all features in (partial R-squared and variance inflation factor values)

### With number of recalled notes as dependent variable.

Table 10:

*Variation inflation factor (VIF) values for model with all features in and number of recalled notes as dependent variable*

| Predictor | VIF |
| --- | --- |
| condition | 1.63 |
| log(attempt) | 1.00 |
| N | 4.59 |
| tonalness | 2.22 |
| i.entropy | 4.50 |
| step.cont.loc.var | 5.46 |
| d.entropy | 1.84 |
| mean\_information\_content | 4.58 |

Table 11:

*Partial R-Squared values for model with all features in and number of recalled notes as dependent variable*

| Effect | F | v1 | v2 | ncp | Rsq | upper.CL | lower.CL |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | 189.72 | 8.00 | 1,758.00 | 1,517.73 | 0.46 | 0.49 | 0.43 |
| log(attempt) | 198.75 | 1.00 | 1,758.00 | 198.75 | 0.10 | 0.13 | 0.08 |
| N | 95.68 | 1.00 | 1,758.00 | 95.68 | 0.05 | 0.07 | 0.03 |
| conditionS | 84.43 | 1.00 | 1,758.00 | 84.43 | 0.05 | 0.07 | 0.03 |
| tonalness | 73.79 | 1.00 | 1,758.00 | 73.79 | 0.04 | 0.06 | 0.02 |
| d.entropy | 62.65 | 1.00 | 1,758.00 | 62.65 | 0.03 | 0.05 | 0.02 |
| step.cont.loc.var | 32.98 | 1.00 | 1,758.00 | 32.98 | 0.02 | 0.03 | 0.01 |
| mean\_information\_content | 0.44 | 1.00 | 1,758.00 | 0.44 | 0.00 | 0.00 | 0.00 |
| i.entropy | 0.04 | 1.00 | 1,758.00 | 0.04 | 0.00 | 0.00 | 0.00 |

### With opti3 as dependent variable.

Table 12:

*Variation inflation factor (VIF) values for model with all features in and opti3 as dependent variable*

| Predictor | VIF |
| --- | --- |
| condition | 1.63 |
| log(attempt) | 1.00 |
| N | 4.59 |
| tonalness | 2.22 |
| i.entropy | 4.50 |
| step.cont.loc.var | 5.46 |
| d.entropy | 1.84 |
| mean\_information\_content | 4.58 |

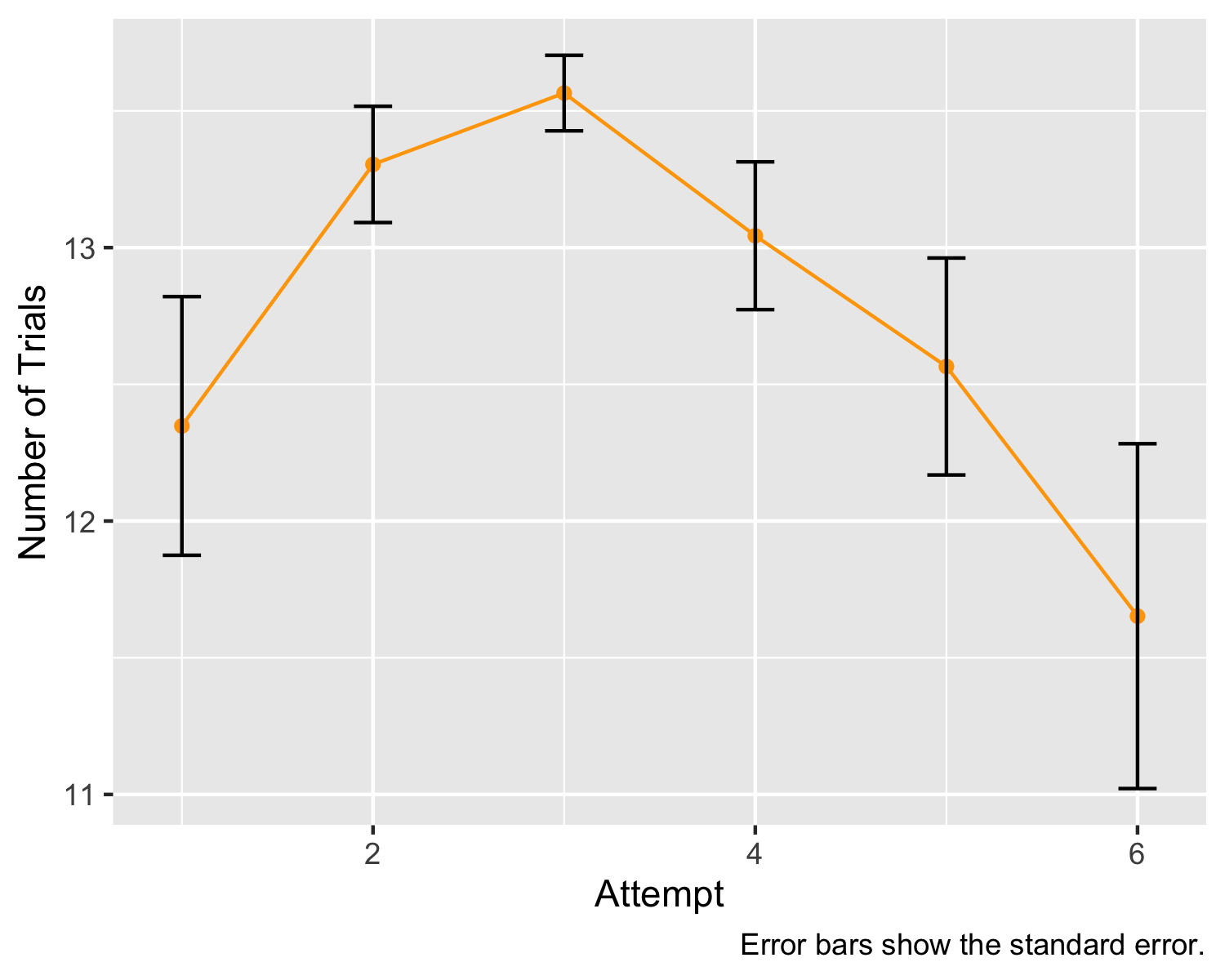
Table 13:

*Partial R-Squared values for model with all features in and opti3 as dependent variable*

| Effect | F | v1 | v2 | ncp | Rsq | upper.CL | lower.CL |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | 42.54 | 8.00 | 1,758.00 | 340.36 | 0.16 | 0.20 | 0.14 |
| mean\_information\_content | 139.38 | 1.00 | 1,758.00 | 139.38 | 0.07 | 0.10 | 0.05 |
| log(attempt) | 86.68 | 1.00 | 1,758.00 | 86.68 | 0.05 | 0.07 | 0.03 |
| i.entropy | 78.48 | 1.00 | 1,758.00 | 78.48 | 0.04 | 0.06 | 0.03 |
| conditionS | 72.03 | 1.00 | 1,758.00 | 72.03 | 0.04 | 0.06 | 0.02 |
| N | 35.68 | 1.00 | 1,758.00 | 35.68 | 0.02 | 0.03 | 0.01 |
| step.cont.loc.var | 2.26 | 1.00 | 1,758.00 | 2.26 | 0.00 | 0.01 | 0.00 |
| d.entropy | 0.44 | 1.00 | 1,758.00 | 0.44 | 0.00 | 0.00 | 0.00 |
| tonalness | 0.00 | 1.00 | 1,758.00 | 0.00 | 0.00 | 0.00 | 0.00 |

# Appendix I

## Counts of overall number of trials that participants utilise for multiple attempts



To assess whether the change across attempts depended on musical experience, we fitted a mixed effects model with trial count as the dependent variable, participant as random effect and the following fixed effects: linear terms for attempt and musical experience; an additional quadratic term for attempt; a linear interaction term for attempt and musical experience; and a quadratic interaction interaction term for musical experience. The model is presented below.

Table 14:

\*\*

| Term |  | 95% CI |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | 11.39 | [10.11, 12.67] | 17.48 | 131.06 | < .001 |
| Attempt | 1.30 | [0.50, 2.11] | 3.19 | 111 | .002 |
| Iattempt^2 | -0.21 | [-0.32, -0.10] | -3.67 | 111 | < .001 |
| Musical experience | -0.67 | [-2.48, 1.15] | -0.72 | 131.06 | .471 |
| Attempt Musical experience | 0.52 | [-0.62, 1.66] | 0.89 | 111 | .376 |
| Iattempt^2 Musical experience | -0.09 | [-0.25, 0.07] | -1.09 | 111 | .278 |

1. <https://www.youtube.com/watch?v=krDxhnaKD7Q> [↑](#footnote-ref-29)
2. <https://github.com/sebsilas/Melodic_Recall_Paper_2023> [↑](#footnote-ref-55)
3. Notenservice Riggenbach. URL: <https://www.notenservice.com/> [↑](#footnote-ref-74)
4. As shown in Appendix F1, however, some participants are closer to approaching the average number of notes in the target melodies by the sixth trial (e.g., *VP24*). [↑](#footnote-ref-89)