Learning and Recalling Melodies: A Computational Investigation Using the Melodic Recall Paradigm

Sebastian Silas & Daniel Müllensiefen
Goldsmiths, University of London, London, United Kingdom & Hochschule für Musik, Theater und Medien, Hannover, Germany

Using melodic recall paradigm data, we describe an algorithmic approach to assessing melodic learning across multiple attempts. In a first simulation experiment, we reason for using similarity measures to assess melodic recall performance over previously utilized accuracy-based measures. In Experiment 2, with up to six attempts per melody, 31 participants sang back 28 melodies (length 15–48 notes) 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 (e.g., tonality, intervallic) aspects of melodies and overall similarity. However, previous melodic recall research mentioned, but did not model, that the length of the sung recalls tends to increase across attempts, alongside overall performance. Consequently, we modeled how the attempt length changes alongside similarity to meet this omission in the literature. In a mediation analysis, we find that a target melody's length, but not other melodic features, is the main predictor of similarity via the attempt length. We conclude that sheer length constraints appear to be the main factor when learning melodies long enough to require several attempts to recall. Analytical features of melodic structure may be more important for shorter melodies, or with stimulus sets that are structurally more diverse than those found in the sample of pop songs used in this study.

Received: November 13, 2022, accepted September 9, 2023.

Key words: melodic memory, melodic similarity, recall memory, melody learning, singing from memory

Most people would not be surprised to hear their friend sing a familiar melody to them, even if their friend was not a great singer or a professional musician. Indeed, remembering melodies is not just an explicitly taught skill useful to professional musicians (Lehmann et al., 2007), but an implicitly acquired ability that most of the general population engage in effortlessly (Bigand & Poulin-Charronnat, 2006; Bigand et al., 2005; Ettlinger et al., 2011; Müllensiefen & Halpern, 2014; Schellenberg et al., 2019; Tillmann et al., 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 as a general skill, in principle 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) and others (Buren et al., 2021; Hallam & Creech, 2010) recently argued, emphasize the importance of musical *production* tasks to gaining a comprehensive understanding of musical abilities. Furthermore, we take modeling of the melodic recall paradigm forward in two main respects. First, we reason for and employ algorithmic similarity metrics to score melodic recall data, noting some limitations of previous approaches, and suggest similarity metrics better help us understand melodic recall processes. Second, while 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 modeling the change in attempt length is a fundamental omission in previous melodic recall studies (e.g., Ogawa et al., 1995; Sloboda & Parker, 1985; Zielinska & Miklaszewski, 1992). In particular, we suggest that modeling the change in attempt length in parallel to the change in overall performance (as measured by melodic similarity metrics) offers at least three main advantages for melodic recall research. First, it points to the

Music Perception, volume 41, issue 2, pp. 77–109, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2023.41.2.77

potential application of models and ideas from already well-established theories of produced mental representations in nonmusical domains (nonmusical 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 behavior may be explained in musicological terms, perhaps suggesting that domain-general memory mechanisms should not be overlooked. Lastly, it enables key insights into how the encoding of melodic information works, which may otherwise be lost in statistical inference that does not consider domain-general memory faculties (see Silas et al., 2022, for a related discussion).

In summary, the work we document here suggests that concepts related to general working memory constraints (e.g., nonmusical serial recall, item length) are important in explaining melodic recall, potentially more so than any other musicological considerations (e.g., interval representations, tonality), at least when the length of the melodies certainly requires multiple attempts to sing back all the notes (e.g., length 15–48 notes), or with pop melodies which are relatively simple to sing. In other words: the "melodic recall" of relatively simple pop melodies 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 recognized the important role of working memory in musical behaviors, 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 general working memory capacities (Talamini et al., 2016, 2017), but note, it is not clear that musical training causally influences general working memory (Silas et al., 2022).

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 specialized 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 specialized systems, quite distinct from general working memory. In our own previous research, we have documented the possible scenarios that 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 musical (e.g., sung recall) performance to someone who has had more music 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 mainly mediated by music-specific processes, or could it be more domain-general processes that turn out to be important? If melodies are long enough to require multiple attempts to sing in full, are musical features beyond length clearly important, compared to the length of a melody alone?

In the nonmusical literature on verbal recall exists some relevant nonmusical analogues to the observations that Sloboda and Parker (1985) made, that the attempt length increases across attempts (Anderson, 1972; Chikhaoui et al., 2009). The so-called list length effect is the finding that recognition performance is superior for items that are part of a short list than for items that were part of a long list (Kinnell & Dennis, 2012). Typically, the literature on verbal memory has used lists of unrelated words as stimuli and asks the participant to recall as many items as they can from memory. Over multiple attempts, Murdock (1960) specifically found that the shape of the learning curve across attempts can be described as an exponential curve with an asymptote equal to the number of items in a target list. However, word lists have different properties to melodies, which presuppose serial recall (i.e., a note order) and embody important structural features within interval and

rhythmic patterns. It is not clear if the unique properties of musical stimuli mean that melodic recall processes are underpinned by fundamentally different processes to their nonmusical analogues. 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 that are not retained in memory or are forgotten easily. Differences that participants do detect are supposed to happen in a musical dimension that 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 et al., 1995; Edworthy, 1985; Massaro et al., 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: First, participant responses are limited to a binary decision (i.e., "identical" vs. "not identical"), possibly with a confidence judgement on an ordered scale. This response format discards 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. Second, 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 the 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 et al., 2017; Harrison et al., 2016) and usage of large-scale musical corpora (e.g., Baker, 2021; Pfleiderer et al., 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 hear 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 operationalization 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; Chenette, 2021) 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's (1985) 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 analyzing 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, as we profile in Experiment 1 of this manuscript.

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 & Hatano, 1988), participants (more participants, participants with and without absolute pitch or formal musical background; Ogawa et al., 1995; Zielinska & 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 music training of the participants. With music students possessing absolute pitch, the point is already at the second trial, whereas music students without absolute pitch needed 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 memorized 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. First, 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 analyzed was limited. For example, Sloboda and Parker (1985) analyzed 48 renditions from their participants, while Oura and Hatano had 320, and Zielnska and Miklaszewski (1992) had 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 that require many data points to be used effectively. Second, previous methods to assess the quality of a sung recall involved accuracy-based measures, which alone are inadequate for meaningfully assessing melodic recall behavior. Third, while 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 the above described shortcomings of previous melodic recall research, we make two general methodological advances: 1) to employ an algorithmic analysis of melodic recall data, and specifically, use similarity metrics; and 2) to model the change in attempt length across attempts in addition to changes in melodic

similarity, which allows comparisons to general memory faculties.

To that end, we conducted two experiments. Experiment 1 is a simulation study where we describe and compare similarity-based metrics to accuracy-based approaches close to those previously taken. Formally, we show how similarity metrics converge and diverge with previous measures for different simulated conditions. This highlights the different properties of similarity-based metrics and how they can be used effectively 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 length of sung recalls changes across consecutive attempts and how this changes alongside overall performance, as measured by melodic similarity.

Experiment 1: Similarity Measures as a Methodological Advance for Scoring Melodic Recall Data

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 basic (i.e., note-by-note) accuracy measures do not meaningfully represent musical structures, such as interval patterns, which involve certain note orders. Hence, such measures of accuracy are also inadequate for scoring sung recall data. For those unfamiliar with the foundations of accuracy assessment (derived in musical context), see Appendix listing #1 for online supplement source material.

Sloboda and Parker (1985) attempted to improve on simple accuracy measures by considering note order in the measure they scored melodic recall data with. They recognized that, in melodic recall data, the target melody and sung 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). Sloboda and Parker's (1985) method of scoring their data was only partly able to address this limitation. They attempted to embody important structural information (interval patterns, or as they called it "contour") into the scoring of pitch data, and additionally, assessed other important (e.g., rhythmic, harmonic) domains separately. However, while their method, and similar methods taken after them (e.g., Koh, 2002; Ogawa et al., 1995), were construed from reasonable musicological considerations, they lack a precise, computational foundation. Second, various domains (pitch, rhythmic, harmonic) were assessed separately, but not combined into a single aggregate measure which weights the domains according to their relevance in human cognition. Third, such comparisons were not tested for their ecological validity (i.e., compared with or based upon human perceptual judgements). Recognizing limitations of their approach at the time, Sloboda and Parker (1985) noted "there is no theory of melodic identity."

MELODIC SIMILARITY

To address this issue, we introduce the notion of a melodic similarity metric for scoring melodic recall data. 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 et al., 2007; Yuan et al., 2020). The similarity measures employed in this study are favored 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 some algorithms that at least came quite close in emulating musically experienced participants' similarity judgements. However, while measures of similarity have been validated in the context of human perceptual judgements (e.g., to predict court case outcomes), we are not aware that they have been profiled in the context of sung recall, as we intend to do here.

Methodological Background

Having obtained numerical representations of both 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 employed here comply with the two main points already raised by Sloboda and Parker (1985) in their 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. Sloboda and Parker (1985) manually attempted a form of alignment, making it a cumbersome task, but additionally, their method is not precisely described. To advance on this point, we detail precise computational approaches to alignment.

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, for example, 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 $[0, 1]$ we use the following:

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

where $|s|$ and $|t|$ denote the element counts of strings s and t respectively, and $d_e(s,t)$ stands for the Edit Distance between strings s and t.

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, importantly, instead of applying the edit distance to raw pitch 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 (τ) contained within the target melody ($f_s(\tau)$) and the comparison melody ($f_s(\tau)$) (see Uitdenbogerd (2002). Formally:

$$u(s,t) = \sum_{\tau \in s_n \cup t_n} |f_s(\tau) - f_t(\tau)| \tag{2}$$

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

$$\sigma(s,t) = 1 - \frac{u(s,t)}{|s| + |t| - 2(n-1)} \tag{3}$$

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 of pitch intervals.

Harmonic similarity is measured by the *harmcore* measure. This measure is based on the chords implied by a melodic sequence, taking pitches and durations (i.e., segmentation) into account. Implied harmonies are computed using the Krumhansl-Schmuckler algorithm (Krumhansl, 1990) and the harmonic progression of the two melodies are compared by computing the number of operations necessary to transform one harmonic progression 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 categorized 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 based on (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):

$$\begin{aligned} opti3 = {} & 0.505 \cdot \text{ngrukkon} + 0.417 \cdot \text{rhythfuzz} \\ & + 0.24 \cdot \text{harmcore} - 0.146 \end{aligned} \tag{4}$$

where we here present the normalized weights, which constrain the values to be [0,1].

Beyond the target or comparison melody lengths being used to normalize the *opti3* score, we note that *opti3* is dependent on the length of the two comparison melodies further in only a "soft" sense, which is particularly relevant to Experiment 2 of this paper, where we use the sung recall attempt length 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 (if notes are missing, not all intervals or rhythms can be reflected, but the implied harmonies can be the same). It should 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). For concise descriptions comparing the similarity measures, see Appendix listing #2 for online supplement source material. Moreover, to build an intuition on how similarity measures may change over attempts, see Appendix listing #3 for online supplement with notated examples of development in sung recall performance and a qualitative description of their change in similarity.

In summary, similarity measures pay attention to musical features that arise from the relationships between pitch and rhythmic values and could be considered more "global" in nature. Conversely, accuracy measures, which count notes or even intervals (bigrams), do not respect the higher order emergent properties of musical features. Consequently, aggregate similarity measures have a greater ability to represent perceptual properties relevant in human cognition and represent a robust step towards computationally representing a notion of melodic identity. 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). As an aid in developing an intuitive understanding of the different properties arising from scoring melodic recall data with accuracy-style vs. similarity measures, see Appendix listing #4 for online supplement source material for simple example comparisons.

MOTIVATION

In Experiment 1, we suggest *opti3* (Equation 4; Müllensiefen & Frieler, 2004a; Müllensiefen & Frieler, 2004b) as a more appropriate measure of melodic recall than previously taken approaches. Specifically, *opti3* aims to address the limitations Sloboda and Parker (1985) noted, by embodying notions of melodic identity that are based on real human perceptual judgements. We profile this measure against our own computational implementation of the approach Sloboda and Parker (1985) took, as well as a similar measure called *percent melodic identity* (*PMI*; Savage et al., 2018). In doing so, we explore how the measures converge and diverge in a quantitative manner, when applied to assessing how well a sung recall matches a target melody.

METHOD

The method taken here is to start with the stimulus set of melodies that we use with real participants in Experiment 2. Starting with this set, we transform the melodies systematically in various ways so that they become less like their originals. We then compare the transformed and original melody versions on several accuracy and similarity measures, and profile how the measures compare as a function of the number of transformations.

Measures of Melodic Accuracy and Similarity

Our reimplementation of Sloboda and Parker's measure is a special case of the "recall" measure of accuracy (see Appendix listing #5 for online supplement source material), which counts the number of correct notes in the sung recall, where "correct" means "contained in the stimulus." Note that this therefore does not penalize participants for missing notes in the stimulus. However, Sloboda and Parker took two additional steps: First, they manually chose the best alignment of the usually shorter sung recall with the usually longer full stimulus. Second, after they aligned the two melodies, they compared the notes between the target and the recall sequentially, one-by-one, meaning that the order of notes is embodied into the scoring. For this reason, we refer to this measure as "aligned ordinal recall." Formally, Sloboda and Parker's aligned ordinal recall (or "contour", as they called it) measure can be expressed as:

$$AOR(t,s) = \sum_{1}^{ij} a\left(t_i, s_j\right) \tag{5}$$

where *AOR* is the aligned ordinal recall function and t and s represent a subset of the target melody and sung recall respectively, and a represents a scoring function to check that aligned notes are equal according to:

$$a = (t,s)\begin{cases} 1, & \text{if } t = s \\ 0, & \text{else} \end{cases}$$

TABLE 1. *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.

where *t* or *s* represents the *ith* or *jth* note in from the AOR function.

Subsequently, the best alignment step can be achieved computationally by taking the maximum score of all alignment possibilities calculated using Equation 5. We note that here we do not re-implement Sloboda and Parker's assessment of other domains (e.g., rhythm), which they did not combine into a single aggregate measure.

The percent of melodic identity (*PMI,* Savage et al., 2018) measure is calculated according to:

$$PMI = 100\left(\frac{ID}{\frac{L_1+L_2}{2}}\right) \tag{6}$$

where ID is the number of aligned pitches that are identical and L_1 and L_2 are the length of each sequence. *PMI* uses its own specific form of automatic alignment: the Needleman and Wunsch (1970) algorithm. This requires automatic alignment penalties to be specified for opening or extending gaps in the alignment. We use the penalties suggested by Savage and Atkinson (2015) and Savage et al. (2018): a gap opening penalty of 12 and a gap extension penalty of 6.

MATERIALS

Fourteen pop songs were used as the basis for the experimental material. Two different melodies were taken from each song to form a stimulus set of 28 melodies in total (9–21 seconds; length = 15–48 notes; see Appendix listing #6 for online supplement source material). 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 generalization 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, the duration values, or both. Subsequently, all results were aggregated across melodies. We describe each sub-experiment succinctly in Table 1 above. For more detailed information, please consult our experiment code and data (i.e., simulated melodies).[1]

[1] https://github.com/sebsilas/Melodic_Recall_Paper_2023 and https://github.com/sebsilas/gensim

DATA ANALYSIS

First, we assess the descriptive statistics for the variables and compare the coefficient of variation (*CV*) between our measures. The *CV* is not the same as the well-known coefficient of determination (R^2). Instead, the *CV* is the standard deviation (*SD*) of a measure divided by its mean. The *CV* is preferred over the *SD* 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 operationalized in Table 2. An ideal property of a measure is that it should increase or decrease monotonically in the context of the experiment manipulations, so we inspect this in particular.

TABLE 2. *Descriptive Statistics for All Simulation Experiment Results, Ordered by the Coefficient of Variation*

Measure	Mean	*SD*	Coefficient of Variation
ngrukkon	0.29	0.30	1.05
harmcore	0.55	0.39	0.72
opti3	0.42	0.25	0.61
aligned_ordinal_recall	0.53	0.26	0.48
rhythfuzz	0.68	0.30	0.44
pmi	0.62	0.23	0.37

Finally, we formally model all experimental data simultaneously. The dependent variable is the accuracy or similarity score; Measure (*opti3*, *aligned ordinal recall*, *PMI*, etc.) and Experiment (1A-IH) are categorical predictors; length of sung recall is a numeric predictor; and additionally the interactions between 1) Measure and length of sung recall and 2) Measure and Experiment were included as predictors. There were 24,426 observations. To facilitate interpretation of the model parameters, sung recall length was standardized before model fitting (note the other parameters are already [0, 1]).

RESULTS

Table 2 presents descriptive statistics, ordered by descending magnitude of the coefficient of variation. The similarity measures, except for *rhythfuzz*, have higher coefficients of variation compared to *PMI* and *aligned ordinal recall*, suggesting that they capture more variance than accuracy measures, at least in the context of our experiments. This is also suggested by the graphs in Figure 1, whereby similarity measures tend to be more affected than *PMI* and aligned ordinal recall as a function of the musical errors we simulated. The other results shown in Figure 1 can be summarized as follows: 1A) adding jitter to duration values causes *rhythfuzz* to degrade as well as *harmcore* (because it offsets the alignment of harmonic progressions), and consequently, *opti3*. *PMI* and *aligned ordinal recall* are unaffected; 1B) note insertions cause all measures to degrade; 1C) note deletions cause all measures to degrade; *aligned*

TABLE 3. *Regression Model Regressing Score Onto Measure, Sung Recall Length and Experiment, Plus Interactions*

Predictor	*b*	95% CI	*t*	*df*	*p*
Intercept	0.60	[0.57, 0.63]	41.57	24016	< .001
Measureharmcore	−0.02	[-0.06, 0.03]	−0.75	24016	.453
Measurengrukkon	−0.28	[-0.32, -0.23]	−12.58	24016	< .001
Measureopti3	−0.18	[-0.22, -0.14]	−8.22	24016	< .001
Measurepmi	−0.11	[-0.15, -0.07]	−5.41	24016	< .001
Measurerhythfuzz	−0.08	[-0.12, -0.04]	−3.83	24016	< .001
Experimentflip out notes	−0.09	[-0.13, -0.05]	−4.33	24016	< .001
Experimentflip out notes and rhythm jitter	−0.09	[-0.12, -0.06]	−5.57	24016	< .001
Experimentinsertions	−0.04	[-0.08, 0.00]	−1.73	24016	.083
Experimentlength mismatch	0.42	[0.38, 0.46]	21.79	24016	< .001
Experimentrhythm jitter	0.44	[0.40, 0.49]	18.82	24016	< .001
Experimentscramble	−0.23	[-0.26, -0.20]	−13.19	24016	< .001
Experimentsubstitutions	0.12	[0.08, 0.15]	5.80	24016	< .001
Sung recall length	−0.16	[-0.24, -0.09]	−4.45	24016	< .001
Measureharmcore × Sung recall length	0.27	[0.17, 0.37]	5.14	24016	< .001
Measurengrukkon × Sung recall length	0.38	[0.28, 0.48]	7.16	24016	< .001
Measureopti3 × Sung recall length	0.33	[0.22, 0.43]	6.14	24016	< .001
Measurepmi × Sung recall length	0.51	[0.40, 0.61]	9.63	24016	< .001
Measurerhythfuzz × Sung recall length	0.38	[0.27, 0.48]	7.12	24016	< .001

Note: Interactions between Experiment and Measure are not displayed.

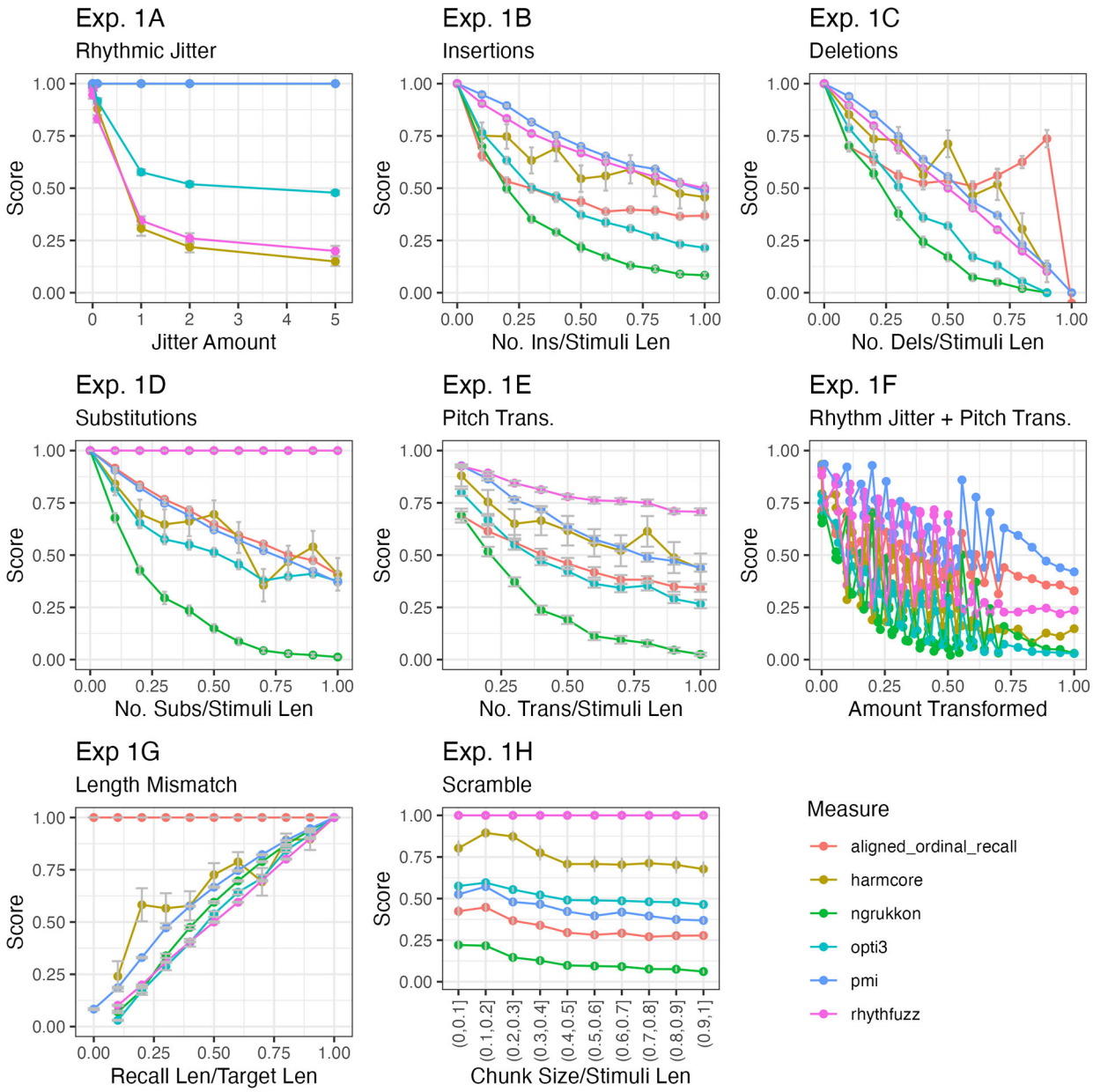

FIGURE 1. Simulation experiment results for accuracy vs. similarity measures.

ordinal recall has a strange property that arises out of the interaction between the alignment process, the recall measure of accuracy, and the ordinal note check, whereby it degrades as the number of deletions approaches half the stimuli lengths, but then increases with further deletions. This nonmonotic property is undesirable; 1D) note substitutions cause all measures to degrade with different rates and function shapes, except *rhythfuzz*, which stays constant; 1E) combined pitch transformations (insertions, deletions, and substitutions) cause all measures to degrade; 1F) combined pitch transformations (insertions, deletions, and substitutions) and rhythmic jitter cause all measures to degrade; 1G) as the length of the recall increases towards the length of the target melody, all measures increase, except *aligned ordinal recall* which is always 1;

1 H) Scrambling the order of pitches does not affect *rhythfuzz*. All other measures deteriorate as a function of scrambling. See Appendix listing #7 for online supplement source material for additional results using basic accuracy and aligned accuracy measures scored on the same data too.

Table 3 shows the general linear regression model results. All measures tend to have lower scores than aligned ordinal recall (the reference level in the regression model), as represented by all negative beta coefficients ($\beta = -0.28$ to $\beta = -0.02$). This corresponds to our descriptive results, whereby the *aligned ordinal recall* and *PMI* measures were less likely to be affected by transformations (simulated human errors), so tend to score more highly. The experiment/transformation type differentially affects the 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 scrambling (Experiment 1 H; $\beta = -.23$) and highest scores rhythmic transformations (Experiment 1A $\beta_{RhythmicJitter} = 0.44$), since most transformations were in the pitch domain. Aligned ordinal recall (the reference level in the regression model), has a negative relationship ($\beta = -.16$) with sung recall length (i.e., scores tend to be lower if a recall is closer to its target in length), possibly representing that, as a sung recall length is longer, there is more "room for error." However, all other measures have positive relationships with sung recall length and a significant interaction with sung recall length. The model had an R^2 value of .44 (adjusted = .44).

DISCUSSION

Our simulation studies generally show that the measures we profiled produce similar assessments of musical behavior, such as sung recall (conversely, see for online supplement source material, which shows that simple accuracy and even aligned accuracy measures produce quite different assessments of musical behavior, and are hence completely inadequate). However, there were some notable limitations of the aligned ordinal recall measure. First, by itself, it is not sensitive to other (e.g., rhythmic, harmonic) errors. Sloboda and Parker (1985) did assess rhythm separately, but they did not formally aggregate it with their aligned recall measure. *PMI* is also not sensitive to rhythmic information. Second, a peculiar and undesirable property arises because of the use of alignment alongside an ordinal recall accuracy check for correctness, producing a nonmonotic effect. This deficiency arises from their being no explicit (or implicit) way for penalizing *misses* in the measure. For this reason, also, third, the *aligned ordinal recall* measure is not sensitive to notes being missed out at the end of a recall, which is what happens when participants cannot yet remember all notes in early trial attempts. Instead, *aligned ordinal recall* has a negative relationship with sung recall length, where all other measures do not. This is a peculiar property, because it suggests that participants can generally score higher by missing notes from the end of their melody; again, an issue with having no implicit way of penalizing for missed notes. The other measures are preferred in this regard because they suggest that, as one sings more notes in the melody, and approaches its true length, they score higher, which is what would be expected. It is not clear if Sloboda and Parker (1985) were aware of this limitation, which was only revealed through simulation. Perhaps it contributes to their lack of an ability to find improvement across trials, which was observed in later melodic recall research (e.g., Koh, 2002).

The melodic similarity measures explicitly embody important musical dimensions beyond pitch and intervals alone and are aggregated together in *opti3*. *opti3* and its weightings are based on human similarity judgements (e.g., Müllensiefen & Frieler, 2007), giving them ecological validity—a property that the other measures do not have. 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 *PMI* and *aligned ordinal recall* do not capture some forms of musical errors, whereas these 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 benevolent on each dimension, but it respects the fact that musical ability is multidimensional: one may sing a rhythm wrong, but the notes right; harmonically the wrong notes 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 melodic 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: 1) The number of different melodies

presented: 14, still not large, but substantially larger to previous research which used 1–2 melodies as targets; 2) The overall number of overall sung recalls to be analyzed (around 2,250); 3) The usage of unambiguously defined and thoroughly tested algorithms of melodic similarity for various musical dimensions; 4) 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 utilize mixed-effect modeling to simultaneously account for the fixed effects of melodic features in explaining participant performance, while also considering participant and item-level random effects, which should ensure that potentially misleading and spurious statistical effects are accounted for; 5) We formally model the change in attempt length (i.e., number of notes recalled). This latter point is based on the observation that Sloboda and Parker (1985) made, that across each new attempt at singing back the same melody, participants tend to contribute more notes than in the previous attempt.

OPERATIONALIZING SIMILARITY VS. ATTEMPT LENGTH

Alongside the similarity measures described in Experiment 1, Experiment 2 introduces the formal modeling of the dependent variable *attempt length*. This represents the number of notes that a participant sings on each trial attempt. Note that this was manipulated in those conditions in Experiment 1 that affected the sung recall length: Experiment 1B (insertions); Experiment 1C (deletions); Experiment 1G (length mismatch). The latter is what specifically simulated the effect of missing out notes from the *end* of a melody, much like has already been observed in melodic recall, on earlier attempts.

Overall, in Experiment 1, sung recall attempt length was associated with lower scores across all *similarity* measures and the *PMI* measure. Hence, the result is intuitive: if you do not sing all notes in a melody, your recall cannot be fully correct; all notes must be present to have sung the melody perfectly. Note, however, the *aligned ordinal recall* 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 is related to similarity, probably causally (numbered of recalled notes => overall similarity), like in Experiment 1G of our simulations, and as we noted earlier: because with fewer notes, some of the melodic identity is destroyed. 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, *attempt length* 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 number of recalled notes via the variable *attempt length*, 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 that consider: 1) relevant experimental conditions (e.g., whether the melody was presented as part of a full audio recording or as melody-only *MIDI* version); 2) features of melodic structure (e.g., melody length, tonality); and 3) 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 does the attempt length submitted change across subsequent attempts 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 and Hatano (1988) in different variants.

Participants

Thirty-one 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, the 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 based on 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. Twenty-three 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 listing #8 for online supplement source material including 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 that 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). Fifteen 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 from 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 stimuli. 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 was piloted informally, but no melodies were discarded.

All songs were used as song excerpts from the original audio recording (audio melodies) and as a 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.[2] 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, for example, 9/8 vs. another 4/4 measure does not affect the results. Likewise, 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 that 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

[2] Notenservice Riggenbach https://www.notenservice.com/

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 earlier 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 listing #9 for online supplement source material) 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 listing #10 for online supplement source material) 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 Attempt Length and Similarity Scores Across Repeated Attempts. To begin our analyses, we inspected our descriptive empirical results. First, we assessed the mean change in attempt length across successive attempts. Next, 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 listing #11 for online supplement source material) across attempt.

Correspondence Between Attempt Length and Melodic Similarity (opti3). Then, changes in the attempt length and in melodic similarity (*opti3*) across attempt were plotted alongside each other on the same graph for comparison. For formal modeling of both attempt length and *opti3*, we proceeded in a mixed effects framework. We constructed two separate models with either a) *attempt length* or b) *opti3* as dependent variable. Consequently, the two models assessed the development of a) the attempt length 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 Modeling. Subsequently, we evaluated a second set of models that additionally included melodic features as predictors. The melodic feature predictors employed were taken from the *FANTASTIC* toolbox (Müllensiefen, 2009) (see Appendix listing #12 for online supplement source material). *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), *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 (Dreyfus et al., 2016; Harrison et al., 2017; Müllensiefen & Halpern, 2014; Silas, Müllensiefen, & Kopiez, 2023). Additionally, to capture the reoccurrence 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 et al., 2020). It is predicted that higher target melody length, interval entropy, and duration entropy, step contour variation, and mean information content will predict worse performance on the task, but higher tonalness will predict higher performance on the task. After iteratively eliminating predictors with a nonsignificant main effect contribution, we tested the interaction between significant feature predictors and attempt. Our feature-based modeling approach is closely related to Baker (2019)'s modeling (who also makes use of features computed from the *FANTASTIC* (Müllensiefen, 2009) toolbox) of melodic encoding and recall processes used in melodic dictation among musicians.

Individual Differences Modeling. Since Sloboda and Parker (1985) dichotomized their participants into "non-musicians" and "musicians," as a means of comparison with their data, we produced graphs of change in dependent variables across attempt, and for illustrative purposes, these were stratified into two groups: high musical experience and low musical experience. These groups were derived by taking the median value of the *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 (attempt length vs. overall similarity) models. We took these values to represent a participant-level latent score reflecting a) the attempt length they can hold in memory b) their overall melodic recall ability on a given attempt. To evaluate whether musical experience is a good predictor of individual differences in both attempt length and overall melodic recall performance, we regressed these participant-level intercepts onto the participant musical experience scores derived earlier. The incremental modeling approach described in the above steps broadly follows the suggestions of Long (2011).

Mediation Analysis. As a means of formally associating attempt length 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, attempt length acted as mediator, and *opti3* as dependent variable.

Correspondence Between Attempt Length and Melodic Similarity (opti3): Revisited. Lastly, we revisited the association between attempt length and melodic similarity by aiming to see if it is generally the beginning or end of attempts that participants focus on improving. Since participants build from an incomplete recall over multiple attempts, we predict that the beginning of the attempt will be better than the later parts of the attempt, because the notes will have been more likely to have been sung in former trials, and hence, be better learned than the later part of the recall.

RESULTS

Assessment of Change in Attempt Length Across Attempt

To visualize the change in attempt length submitted across trials, Figure 2 graphs the mean attempt length, as a function of attempt. As shown, the attempt length increases across successive attempts. The effect is clearly nonlinear, with a diminishing gain in attempt length across attempts. Note that the average target melody length is 25.39. Consequently, even after six attempts, on average, participants are still not submitting close to the number of notes in a target melody.[3]

In the formal mixed effects model, attempt length 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 R^2 value of the mixed effects model was 0.14 and the conditional R^2 value 0.65 (Nakagawa & Schielzeth,

[3] As shown in the online supplemental material, however, some participants are closer to approaching the average number of notes in the target melodies by the sixth trial (e.g., *VP24*).

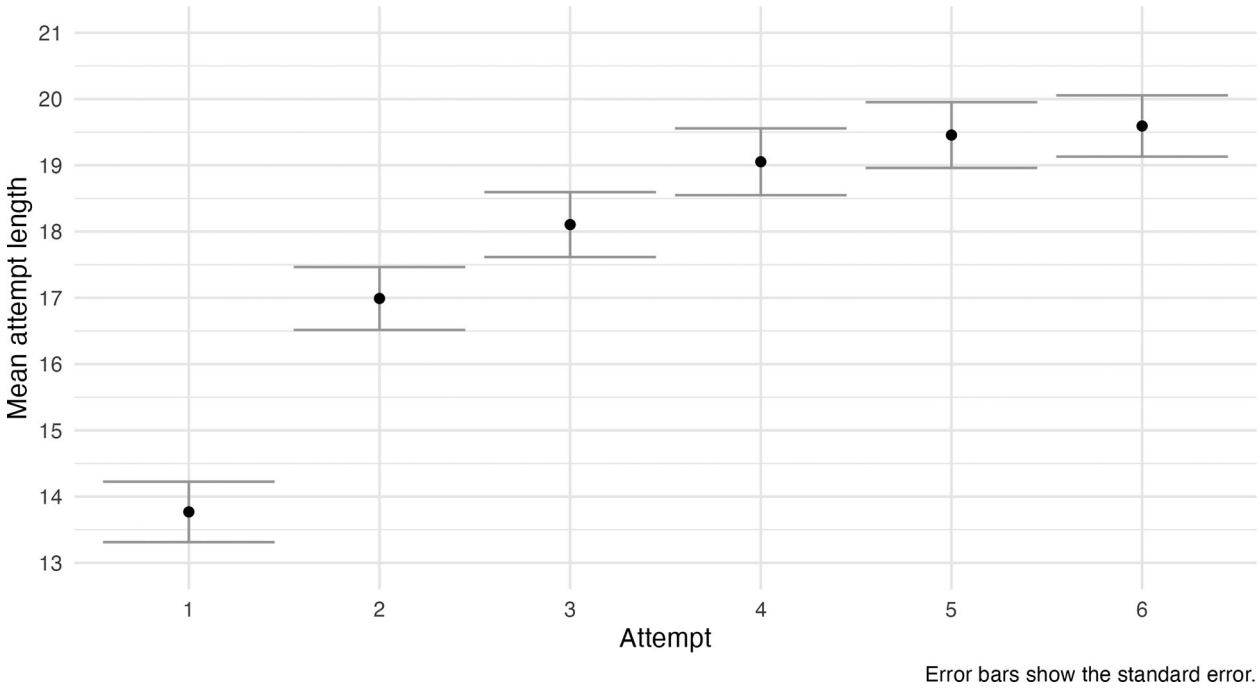

FIGURE 2. Development of average attempt length across attempt.

2013).[4] This suggests that the fixed effects (attempt and condition), while 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 R^2 value (to 0.79) and the marginal R^2 value slightly (to .143). This final model (Model *A1.2*) is shown in Table 4.

TABLE 4. *Model A1.2: Mixed Effects Model Regressing Attempt Length Onto Attempt and Condition*

Term	$\hat{\beta}$	95% CI	*t*	*df*	*p*
Intercept	11.39	[8.15, 14.63]	6.89	32.26	< .001
ConditionS	5.30	[0.99, 9.60]	2.41	25.65	.023
Logattempt numeric	3.71	[3.38, 4.03]	22.09	1,470.74	< .001

Note: The index 'S' refers to the 'Sound' condition.

Assessment of Similarity Scores Across Repeated Attempts
To visualize higher level changes in melodic recall performance (indicated by similarity) across attempts, Figure 3 graphs the mean score of each similarity measure (*opti3*, *ngrukkon*, *harmcore* and *rhythfuzz*) as a function of attempt. A linear model (represented by solid-colored lines) would suggest a general increase across attempt for all variables, except *harmcore*, which appears relatively stable across attempt (see Appendix listing #13 for online supplement source material). However, as seen with attempt length, while a linear model predicts the data over the course of six trials reasonably well, for the generally increasing variables, (*opti3*, *ngrukkon*, and *rhythfuzz*) a nonlinear 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, $B = .07$, $p < .001$; condition, $B = .10$ ($p = .01$). 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 R^2 of .098 and a conditional R^2 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 R^2 (to .101) and the conditional R^2 (to .71). See Table 5 for the final model (Model B1.2).

[4] Note, the marginal R^2 represents the variance explained by the fixed effects only; the conditional R^2 represents all variance explained by the model (fixed and random effects).

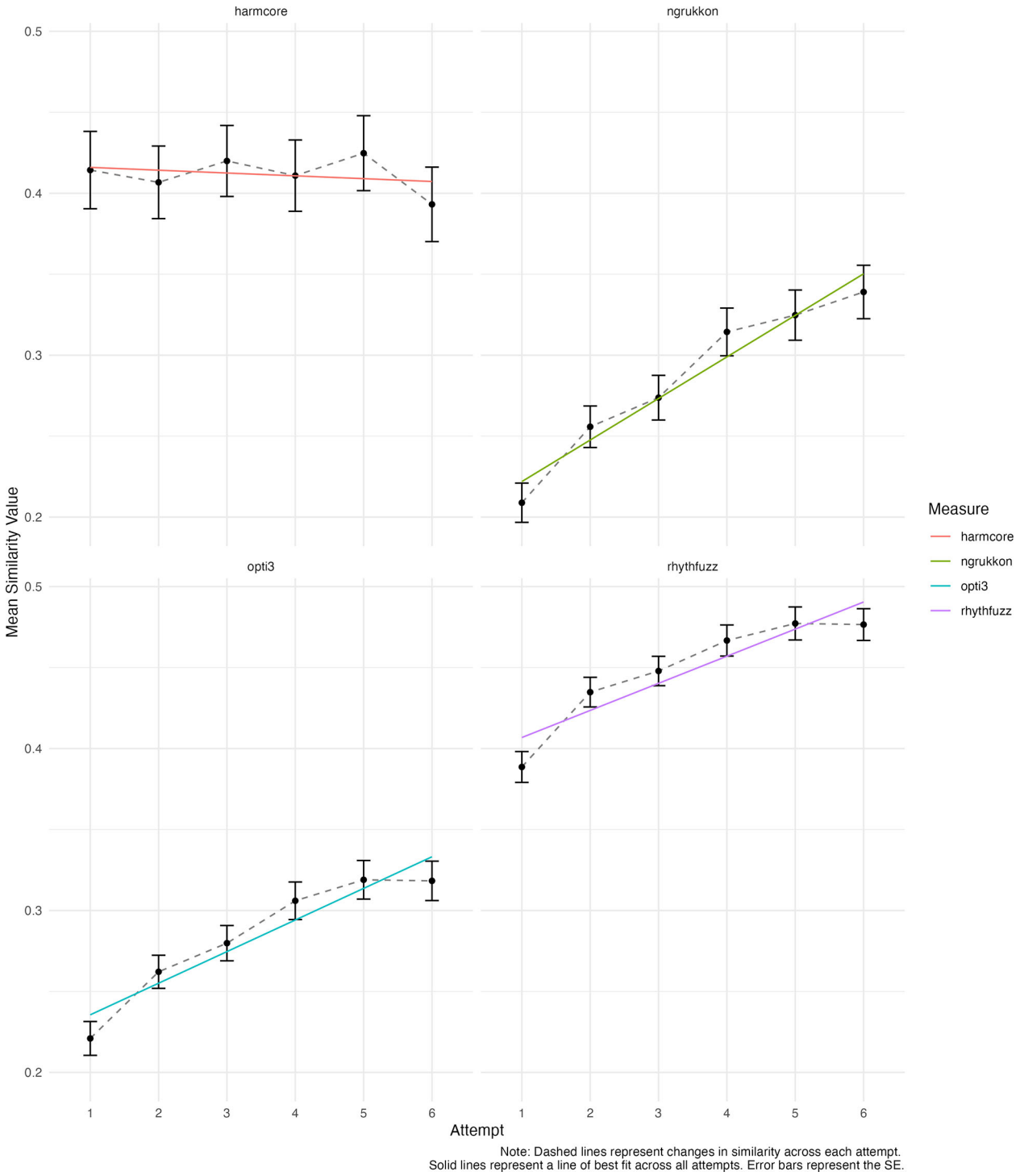

FIGURE 3. Mean similarity values as a function of attempt.

TABLE 5. *Model B1.2: Mixed Effects Model Regressing the Similarity of Melodic Recalls (opti3) Onto Attempt and Condition*

Term	$\hat{\beta}$	95% CI	*t*	*df*	*p*
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 numeric	0.07	[0.06, 0.08]	15.88	1,460.86	< .001

Correspondence Between Attempt Length and Melodic Similarity (opti3)

The development of both overall similarity (*opti3*) and the attempt length across attempt is broadly similar in shape: increasing, with diminishing gains on each attempt. To make this clear, we rescale the attempt length variable to be in the range 0 to 1, like *opti3* and plot them alongside each other in Figure 4.

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 attempt length alongside the overall change in melodic recall performance (here indicated by melodic similarity) across recall attempts. Consequently, we proceed by modeling the two effects via two sets of mixed effects models in parallel. We take forward both models *A1.2* and *B2.2* (Tables 4 and 5), where 1) condition and log attempt are always included as fixed effects and 2) 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 attempt length and *opti3* converge, but diverge, and hence, measure different constructs. The bivariate linear Pearson's 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 Modeling

For modeling the memorability of melodies we added *target 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 R^2 increased significantly from .14 to .46 when *attempt length* was dependent variable and marginally from .10 to .16 when *opti3* was dependent variable. However, after removing nonsignificant predictors, only *target melody length* was

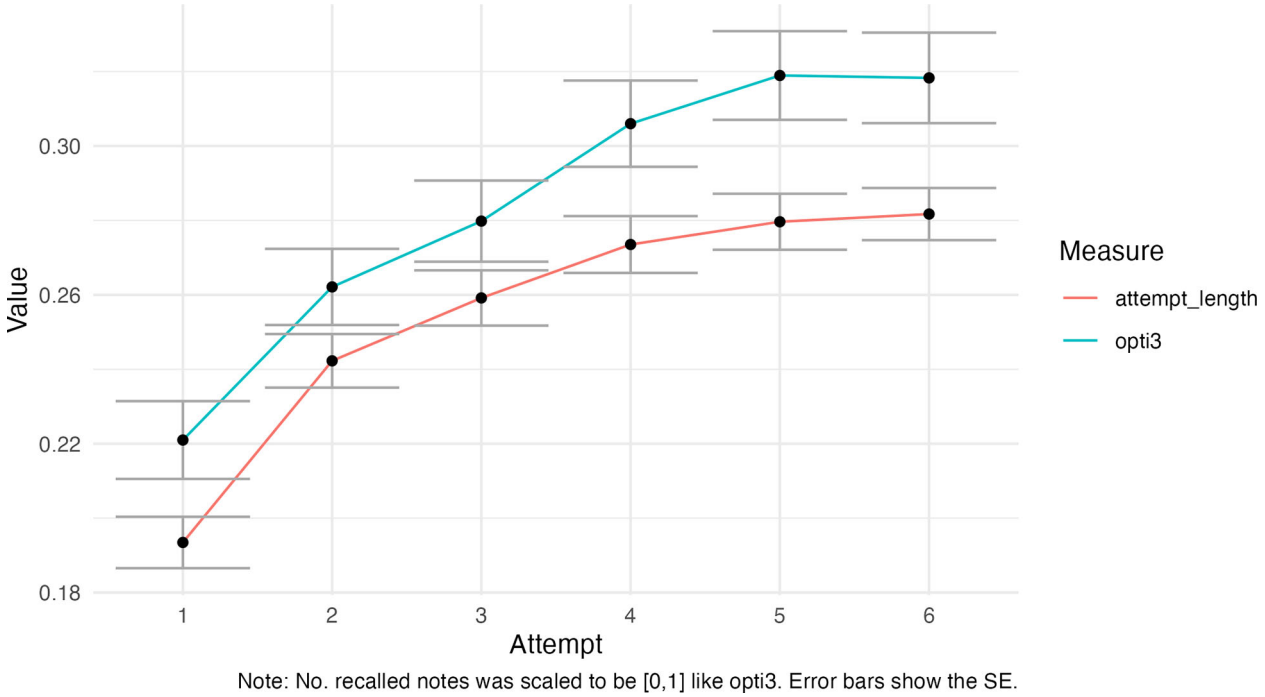

FIGURE 4. Development of mean no. recalled notes (red) and opti3 (blue).

a significant predictor when *attempt length* was dependent variable, and none were significant with *opti3* as dependent variable. In a model with *attempt length* as dependent variable, and only *target melody length* as fixed effect predictor alongside log attempt and condition, the marginal R^2 was .44, suggesting that the other melodic feature predictors really do not add much explanatory power to the model. We also present variance inflation factors and partial R^2 values for diagnostics (see Appendix listing #14 for online supplement source material). Altogether, the interpretation that the nonsignificance of the other melodic features is due to high collinearity can be ruled out, and it is evident that *target melody length* substantially explains variance in *attempt length* by itself. Note also, 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 (see Appendix listing #15 for online supplement source material). This also suggests that melodic features have sufficient heterogeneity beyond melody length, which empirically has *less* heterogeneity.

We tested the interaction between *attempt* and *target melody length* in the model with *attempt length* as dependent variable. The interaction term was statistically significant ($B = 0.21$, $p < .001$), suggesting that the length of the melody differentially affects the attempt length, depending on the attempt number. For this final model (A2.2), the marginal R^2 value was .45 and the conditional R^2 value .80.

Individual Differences Modeling

Individual Differences in Changes of Attempt Length Across Attempt. Figure 5 presents changes in attempt length 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 nonlinear increase in attempt length across attempt for both groups. However, while both groups submit a similar attempt length to the first attempt on average, the higher musical experience group tend to submit more notes to each subsequent attempt (however, for an alternative by-participant comparison see Appendix listing #16 for online supplement source material). 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 steeper. This

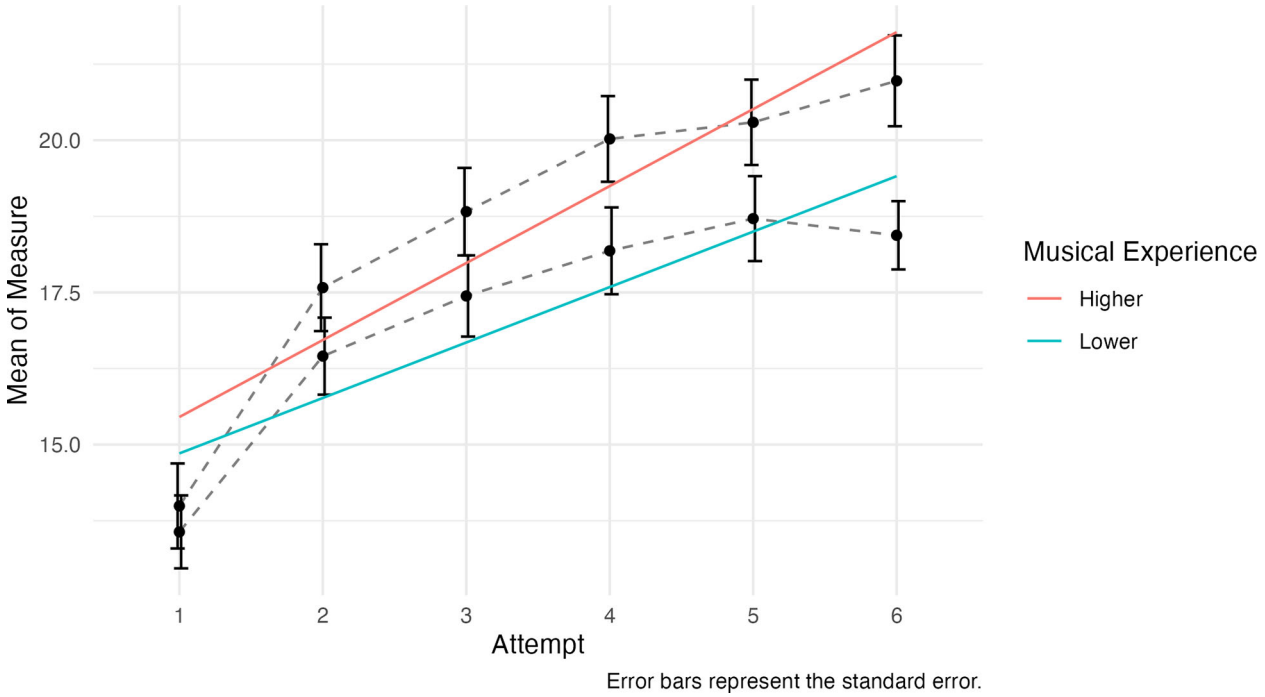

FIGURE 5. Development of average attempt length across attempt, factored by level of musical experience.

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 dichotomization of musical experience formally.

Individual Differences in Similarity Changes Across Attempt. The corresponding figure for similarity measures (*opti3* et al.), Figure 6, 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 similarity of their recalls more effectively than participants with lower musical experience. The difference in slopes is also notable for overall similarity (*opti3*). For an alternative by-participant visualization, see Appendix listing #16 for online supplement source material.

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 attempt length, 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 *attempt length* 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 R^2 value of .36 (adjusted = .33), $p < .01$ and musical experience as the sole significant predictor, $B = 0.05$, $p < .01$. A similar pattern was seen for the model built with *opti3* as dependent variable: only *musical experience* was a significant predictor ($B = 1.63$, $p < .01$). This model had a small R^2 value of .24 (adjusted = .20).

Mediation Analysis

Earlier, when adding melodic features as fixed effects to the base mixed effects models described above, we found that only *target length* was a statistically significant predictor of performance when *attempt length* was dependent variable and none when *opti3* was dependent variable. However, as noted previously, there is a correspondence between attempt length and overall similarity across attempts (see Figure 4). Perhaps target length could indeed predict overall performance (*opti3*) *via* an effect on attempt length, a hypothesis that can be implemented as a mediation model. We tested this hypothesis using the *mediate* function from the *R* package *mediation* (*v* 4.5.0). *Target Length* was treated as main fixed effects predictor, *opti3* was the dependent variable, and *attempt length* was the mediator. The input to the function is two nested multiple regression models which leave all significant predictors, including interactions, from earlier steps to make sure all already discovered effects are simultaneously modeled and accounted for. These effects were specified as covariates with respect to Target Length. In our case, the two nested regression models are:

Fit Mediator.

$$\begin{aligned} AttemptLength = {} & Condition + TargetLength \\ & + (\log(Attempt) * TargetLength) \\ & + ParticipantByItem{*}\log(Attempt) \end{aligned} \tag{7}$$

where *Condition*, *TargetLength* and the interaction between log(*Attempt*) and *TargetLength* are fixed effects and log(*Attempt*) has a random slope for each participant-item grouping, *ParticipantByItem*.

Fit Dependent Variable.

$$\begin{aligned} opti3 = {} & Condition + AttemptLength + TargetLength \\ & + (\log(Attempt){*}TargetLength) \\ & + ParticipantByItem{*}\log(Attempt) \end{aligned} \tag{8}$$

where *Condition*, *AttemptLength*, *TargetLength* and the interaction between log(*Attempt*) and *TargetLength* are fixed effects and log(*Attempt*) has a random slope for each participant-item grouping, *ParticipantByItem*.

Note that, because we found there is no direct relationship between the dependent variable (*opti3*) and the independent variable (*Target Length*) *a priori*, our use of mediation is known as inconsistent mediation (Hayes, 2009; MacKinnon et al., 2007). It was predicted that *Target Length* will have a negative direct relationship with *opti3*, since longer melodies should place more of a strain on working memory and contribute to a worse performance. See Figure 7 for a representation of the mediation model.

In the mediation model results, the Average Direct Effect $B = -0.005$, $p < .001$ and the Average Causal Mediation Effect $B = -0.003$, $p < .001$ were statistically significant (note that the *Average Causal Mediation Effect* can be manually derived by multiplying the path coefficients from *Target Length* to *Attempt Length* and *Attempt Length* to *opti3* (i.e., 0.32 * 0.01 = 0.003). Consequently, the *Total Effect* $B = -0.001$, $p = .40$ was not

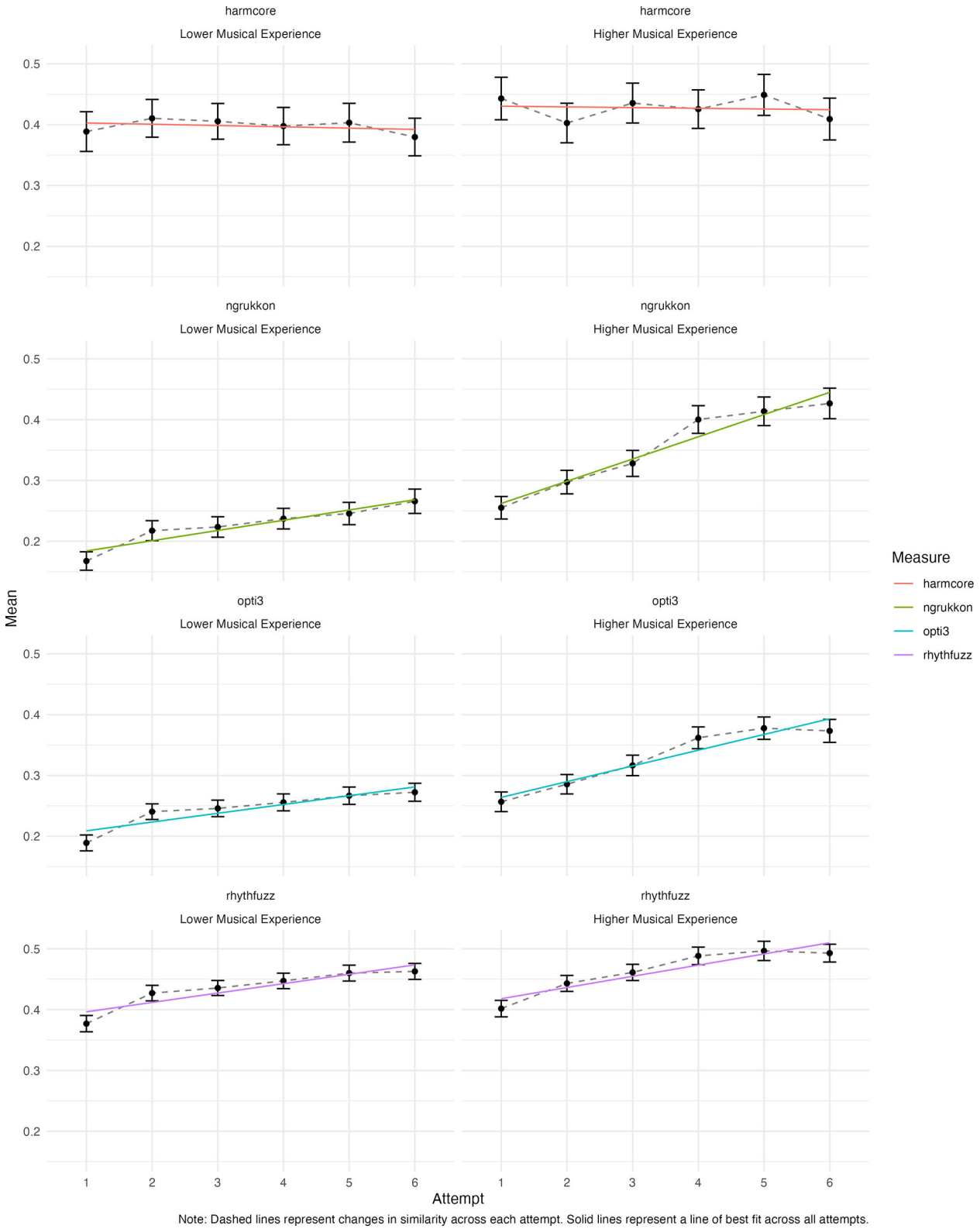

FIGURE 6. Similarity scores as a function of attempt, dichotomized on musical experience.

FIGURE 7. Mediation model results whereby Attempt Length acts as mediator between Target Melody Length and Melodic Similarity (opti3).

TABLE 6. *Model Comparison With and Without Attempt Length as Fixed Effects Predictor*

	npar	AIC	BIC	logLik	Chisq	Df	Pr(>Chisq)
fit.dv.excl.attemptlength	9	−1996	−1947	1007			
fit.dv.incl.attemptlength	10	−2274	−2220	1147	281	1	$p < .001$

Note: The two models compared correspond to the Fit Dependent Variable model given in Eq. 8. The two models differ only in the presence/absence of AttempLength as a fixed effect in the model equation.

statistically significant, since the direct and indirect effects are two opposite effects working against each other, producing a nonsignificant effect at the level of the bivariate relationship between Target Length and *opti3* (known as inconsistent mediation; Hayes, 2009; MacKinnon et al., 2007). This suggests that, overall, longer melodies lead to longer attempt lengths (because the target melody is longer, so requires more notes), which in turn is associated with increases in the *opti3* score, presumably because longer attempt lengths also generally contain more improvements along the domains of harmonic, rhythmic, and intervallic information (reflecting learning from previous attempts). However, longer target melody length contributes to a melody item's melodic complexity and difficulty (or is a proxy for general complexity), and hence, the ability for a given melody to be held in working memory, explaining the negative direct relationship with opti3.

That the Average Causal Mediation Effect was statistically significant in both models suggests that target melody length can indeed be a predictor of overall performance (as indicated by *opti3*), at least partly via its influence on attempt length submitted to an attempt. In order to test whether including attempt length as fixed effect predictor was justified, we compared two versions of the Dependent Variable Model (Eq. 8), with and without attempt length as predictor: the model with attempt length had a lower BIC value (-2274) and hence the better fit than the one with (BIC = -1947). See Table 6.

Modeling the Correspondence Between opti3 *and* attempt length *Revisited*

Sloboda and Parker (1985) observed that the sung recalls got considerably longer over six attempts, but the ratio between the number of correctly recalled notes and the overall number of sung notes stayed approximately constant over trials. Consequently, the *number* of errors increased across attempts. This observation suggests the following cognitive processes may be taking place: 1) Participants add more new notes on each attempt, since they have remembered some, but not all, of the target melody from the previous attempt. 2) The parts of the melody they have attempted to recall from the previous attempt should be more likely to improve on subsequent attempts. It seems most likely that participants attempt to recall the beginning of the melody first and gradually add more notes to the end: a kind of primacy bias. However, it may also be that participants exhibit a recency effect and (or) attempt to successfully recall the end of the melody. Alternatively, perhaps there is no bias at all, and participants just improvise a gist of the whole melody or improve on different parts somewhat randomly. To investigate this question with our data, Figure 8 visualizes changes in similarity as a function of the sung recall section (beginning, middle, and end) and attempt number. All recalls were divided into three equal sections. For all recalls to be divided into three truly equal sections (i.e., not always leaving one section unbalanced where equal division is not possible,

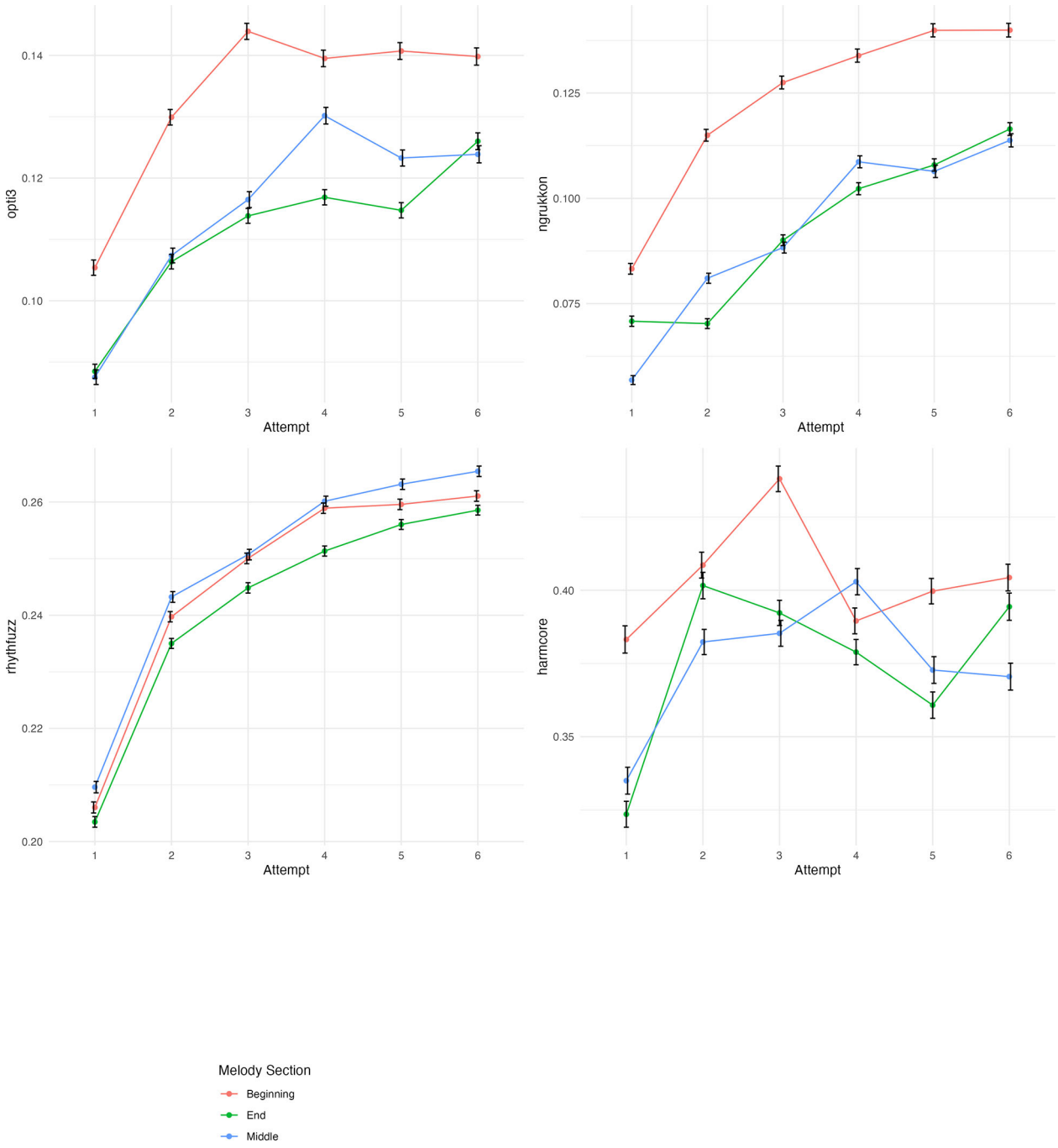

FIGURE 8. Changes in similarity as a function of attempt and sung recall section (beginning, middle, end).

which might produce a bias in our scoring), the separation of the recall into three distinct sections is not always contiguous, and hence, some of the same notes may appear in two sections.

Figure 8 suggests the following patterns: 1) The overall similarity improves across all melody sections across attempts. However, the beginning of the melody is always better recalled, then the middle, then the end. 2) The same effect is seen specifically for interval patterns (*ngrukkon*), with a more nuanced result too: the recall of the beginning of the melody seems to exhibit nonlinear, diminishing gains, across attempts, whereas

improvements in the middle and end of the melody seem to be better represented by linear gains, and that are similar (i.e., statistically indistinguishable), in terms of their level of similarity. 3) Curiously, for rhythmic similarity, the ordering is flipped for beginning and middle, such that the rhythmic performance is always better for the *middle*, rather than the end of the recall. 4) There is a slight bias towards harmonic performance in early attempt (1 and 2), but then there appears to be no difference between level of similarity for melody section for later attempts. For statistical models to support these interpretations, see Appendix listing #17 for online supplement source material.

Discussion

HOW DO WE LEARN TO RECALL MELODIES?

To understand how melodies are learned to be recalled across multiple attempts, we studied how both the attempt length and overall similarity change across the time course of six attempts. Our data suggests that the attempt length starts from an incomplete recall in attempt number one, and in each subsequent attempt, participants generally add more notes. The length of the recall across attempts grows shaped like an exponential curve, but which, on average, asymptotes at the number of notes in the target melody, or less, with six attempts potentially not being enough attempts to successfully recall the full length of the target melody. These results are similar to those presented in the nonmusical 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, 1960). The exact shape of this curve depends primarily on the number of notes in the target melody and the participant recalling the melody. Generally, similarity between the target melody and the sung recall increases across attempts too, suggesting incremental learning of melodic identity across repetitions. This is the case for overall composite melodic similarity (*opti3*) and constituent parts of overall similarity: 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 their recalls, participants tend to focus on improving the earlier part of the target melody correctly, performing less well on the later notes they have added more recently. However, this is not the case for rhythm, where, curiously, the middle of the melody is performed better than the earlier or later parts. Perhaps this suggests that participants primarily focus on pitch/intervallic information, then when they have recalled as many intervals as they can on a given attempt, switch processes to retrieving rhythmic information. In other words, perhaps a domain-specific working memory load for intervallic information has been reached, so participants switch to a less-burdened rhythmic working memory capacity to yield more overall gains. This would resonate with the idea of domain-specific interference in working memory (Jarrold et al., 2011).

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. While the melodies 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 naturally structurally varied to the same degree as intervals and rhythm. Since our approach was to use melodies from real popular music songs, without artificial manipulation, consequently, this simply implies that because Western pop music generally contains melodies in a single key, they naturally contain little tonal variance. Conversely, intervallic and rhythmic structures naturally have more variance. Hence, without artificially manipulating the harmonic structure of melodies to contain more tonal variance (changing key/tonal center), we can only comment on the statistical regularities of melodies that arise in popular music, and their incidental associations with memorability. To establish whether tonality is extracted more easily than other features, beyond its naturalistic variance, further research would need to use melodies which more clearly change key. Then, we suspect that we would 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 mind should implicitly extract such co-occurring statistical regularities (Pearce, 2018). Whether or not this interpretation is true, we highlight our finding that improvement *can* be indicated across trials, which is *not* in accordance with Sloboda and Parker (1985), who only observed that attempts got longer, but not better. That both attempt length and performance increase across trials has been observed in other research too (e.g., see Koh, 2002).

With respect to individual differences, first, 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 speaking, this could be because some participants have previously implicitly learned similar melodies to those they were tested with. Alternatively, perhaps some melodies contain features that rely more on musical vs. nonmusical memory than others: the former might benefit highly musically experienced participants, and the latter, those with very good nonmusical 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. This is also indicated by musical experience being a significant predictor of both attempt length and overall similarity. However, this is not the case for all participants (see Appendix listing #16 for online supplement source material): some participants low on musical experience can still learn quickly across trials, presumably because they can nonetheless make large improvements over trials as a function of other abilities, such as their general working memory. As explored in Silas et al. (2022), high general working memory capacity 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.

In our data, and as Sloboda and Parker (1985) previously noted, it seemed that a dominant factor in performance is the attempt length. This effect might be more to do with the sheer need for rote repetitions to remember a target for learning, rather than being mainly about utilising musical structures. In this way, each attempt is a new iteration adding more note events to the long-term memory store for a particular melody. While we would expect the extraction of musical features to somewhat mitigate general capacity limits, because structure helps memorability (Gobet, 2005; Gobet et al., 2001; Müllensiefen & Halpern, 2014; Thalmann et al., 2019), perhaps it is not so surprising that the sheer attempt length might be such an important variable. With respect to general theories of working memory capacity constraints, the lengths of the melodies we used were relatively long (*target melody length* = 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 on working memory could be broadly responsible for the producing the increasing length of recall across attempts, as we and Sloboda and Parker (1985) observed (Cowan, 2010; Miller, 1956; Shiffrin & Nosofsky, 1994; Vergauwe et al., 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, & Kopiez, 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 modeling, beyond musical features and musical memory (Silas et al., 2022).

However, participants do recall more notes than Miller (1956)'s 7 +/- 2 rule would suggest. On the first attempt in our melodic recall data, participants are recalling nearly double this number of notes on average (Figure 2). This suggests that musical structures may be playing a role, but since the similarity is still low, participants are improvising some extra notes in their attempts, even though these extra notes are relatively incorrect. Nonetheless, previous research has observed a role for musical features in predicting memorability. For example, Cuddy et al. (1981) showed that chromatic melodies are harder to learn than diatonic ones and Povel and Essens (1985) found that more complex rhythms are harder to learn that simpler ones. Perhaps if we used more complicated melodies, which do not

come from pop songs, we would find that melodic features beyond target melody length can be connected to recall performance. As noted, we have been able to successfully do this in other research (Silas, Müllensiefen, et al., 2023), where the corpus of melodies used was more varied in terms of their musical features, and where we had a much larger item bank of melodies, producing more variance. However, we used much shorter melodies there, which we think is what mainly contributes to the explanatory significance of melodic features in that study (i.e., melodic features are important if the melody is not too long to recall as to require multiple attempts). Nonetheless, it is important to note that perhaps the failure to connect melodic features to performance in the present study represents something idiosyncratic about the stimulus set.

WHAT MAKES MELODIES DIFFICULT TO REMEMBER?

Consequently, when including variables to indicate melodic complexity as predictors in our mixed effects model, we found that melodic features were not statistically significant predictors of melodic recall performance with *opti3* as dependent variable. As noted, however, the attempt length seemed to be a main factor that might be dominating overall melodic recall performance—and target melody length was a significant predictor of attempt length. That our measure of similarity, *opti3*, also increases across trials indicated that similarity might increase predominantly as a function of the attempt length (as indicated also by our simulation experiment 1G). Taking this observation into account and including *attempt length* as a mediator between target melody length and *opti3* as dependent variable, connected target melody length to *opti3*, 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 attempt length. 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 (hence the negative direct relationship between *target melody length* and *opti3*).

Nonmusical models of serial recall predict similar effects to those seen in our data. For example, as Anderson (1972) notes, Murdock (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, while musical features could make melodies more or less difficult to remember, perhaps the main difficulty across multiple trial attempts is the sheer length of the target melody itself, and the current working memory load (Baddeley & Hitch, 1974), at least: 1) with the current melody set of relatively simple pop melodies and 2) when the length of melodies is 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 attempt length should be integratively modeled. In this paper, we did this via mediation modeling. In further research, more detailed relationships including other variables (e.g., music 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 elicitation of verbal memory and social psychological systems (Clayton, 2008; Tarr et al., 2014).

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 target melody, should it require multiple attempts to sing all the notes back. On each attempt, more notes will be recalled, and the attempt length will approach the number of notes in the target melody (similar to models of nonmusical serial recall; Anderson, 1972), or a long-term memory constraint dependent on the timespan of learning, over the time course of several trials. But, so long as the participant adds new notes and improves on the earlier parts of their recall in subsequent attempts, remembering better the melodic structure they already tried to recall in previous attempts, and the attempt length 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 help to learn

melodies quickly, presumably because of mental templates that 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 expertise 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 attempt length 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 attempt length across attempts are causally responsible for improvement in *opti3* across attempts. For instance, it could be that *opti3* increases across attempt alongside *attempt length* simply in an associative manner, whereby *opti3* increases despite the associated increases in *attempt length*. However, beyond the strong associative pattern, there are strong logical and inductive grounds for supposing the causality. Most importantly, as noted, 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 *attempt length*. 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 *attempt length* and *opti3* should be even more integratively modeled, using a much larger database of items and more heterogeneity in melodic features. We have recently implemented such a framework (Silas, Müllensiefen, & Kopiez, 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 or hierarchical structure, 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 abilities should be simultaneously measured to understand and represent both domains properly. Lastly, since effects around item length are modeled and described well in the *ACT-R* framework, which has several models of serial recall (e.g., Anderson, 1972) relevant to the attempt length, and emphasizes modeling produced events, we intend to more thoroughly explore modeling that integrates the *ACT-R* framework (Ritter et al., 2019) alongside melodic feature modeling. 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 modeling framework that is primarily concerned with explaining musical production seems highly relevant to investigate musical abilities in a comprehensive way and beyond perceptual paradigms in the future (Okada & Slevc, 2021).

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, Niels Verosky, and David Temperley 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

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* [Doctoral dissertation, Louisiana State University]. 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 S. Hallam, I. Cross, & M. H. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 35–44). Oxford Academic. 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

Cuddy, L. L., Cohen, A. J., & Mewhort, D. J. K. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 869–883. https://doi.org/10.1037/0096-1523.7.4.869

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception and 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*, *3*(3), 259–274. https://doi.org/10.2307/40285336

Dowling, W. J. (1972). Recognition of melodic transformations: Inversion, retrograde, and retrograde inversion. *Perception and 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 and Psychophysics*, *50*(4), 305–313. https://doi.org/10.3758/bf03212222

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

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 and Psychophysics*, *57*(2), 136–149. https://doi.org/10.3758/bf03206500

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*, *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 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

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, *76*(4), 408–420. https://doi.org/10.1080/03637750903310360

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 and Psychophysics*, *24*(6), 551–565. https://doi.org/10.3758/BF03198783

Jarrold, C., Tam, H., Baddeley, A. D., & Harvey, C. E. (2011). How does processing affect storage in working memory tasks? Evidence for both domain-general and domain-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 688–705. https://doi.org/10.1037/a0022527

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

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory and Cognition*, *40*(3), 311–325. https://doi.org/10.3758/s13421-011-0164-2

Koh, C. K. (2002). *Memory and learning in music reproduction: The effects of melodic structure, perceptual cues and learning methods on music recall* [Doctoral dissertation, Queen's University].

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

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

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

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

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 (Eds.), *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). International Conference of Music Perception and Cognition.

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. (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 I. Deliège & J. Davidson (Eds.), *Music and the mind: Essays in honour of John Sloboda* (pp. 161–186). 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

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

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., & 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

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). Springer. https://doi.org/10.1007/978-3-642-11674-2_16

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

Povel, D.-J., & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*(4), 411–440. https://doi.org/10.2307/40285311

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). 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

Savage, P., & Atkinson, Q. (2015, January 1). *Automatic tune family identification by musical sequence alignment*. 16th International Society for Music Information Retrieval Conference.

Savage, P., Cronin, C., Müllensiefen, D., & Atkinson, Q. (2018, June 26). *Quantitative evaluation of music copyright infringement*. Quantitative evaluation of music copyright infringement.

Schellenberg, E. G., Weiss, M. W., Peng, C., & Alam, S. (2019). Fine-grained implicit memory for key and tempo. *Music and 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 a singing test based on item response theory and a general open-source software environment for singing data. *Behaviour Research Methods*. https://doi.org/10.3758/s13428-023-02188-0

Silas, S., Robinson, M., Baker, D., Müllensiefen, D., Harrison, P., & Jacoby, N. (2023). *The recall-recognition paradox? A multi-paradigm approach to exploring the memorability of sonic logos*. Manuscript in preparation.

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). 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*, *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* [Doctoral dissertation, RMIT University]. 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

Appendix

Online Supplement Sources

1. Foundations of accuracy assessment (derived in musical context)
Section 1:
https://sebsilas.github.io/musicassessr/articles/accuracy_vs_similarity_measures.html#foundations-of-accuracy-style-measures

2. Descriptions comparing the similarity measures used in this study
https://sebsilas.github.io/musicassessr/articles/musicassessr_dependent_variables.html

3. Notated examples of development of sung recall performance over multiple attempts and a qualitative description of their change in similarity
https://sebsilas.github.io/musicassessr/articles/intuitive_similarity.html

4. Example comparisons for scoring melodic recall data with accuracy-style vs. similarity measures
https://sebsilas.github.io/musicassessr/articles/accuracy_vs_similarity_measures.html#example-comparisons-of-accuracy-vs—similarity-measures-on-the-same-data

5. A comparison of accuracy vs. similarity measures
Section 1: https://sebsilas.github.io/musicassessr/articles/accuracy_vs_similarity_measures.html

6. A sample of the stimulus set used in this study
Section 1:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#short-melodic-excerpts-from-pop-songs-used-as-materials-in-the-study

7. Simulation results of Experiment 1 using basic accuracy and aligned accuracy measures
https://sebsilas.github.io/musicassessr/articles/profile_basic_accuracy_measures.html

8. List of pop songs used as materials in this study
Sections 1 and 2:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html

9. The musical experience questionnaire used in this study
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#questionnaire-items

10. The single factor exploratory factor analysis solution of the musical experience questionnaire
Section 3.2:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#factor-loadings-for-mixed-type-variables-based-on-questionnaire-items

11. Dependent variables used in this study
https://sebsilas.github.io/musicassessr/articles/musicassessr_dependent_variables.html

12. Melodic feature predictors used in this study
https://sebsilas.github.io/musicassessr/articles/melodic_features.html

13. Linear model details for comparison with the nonlinear models taken forward
Section 6:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#linear-vs—non-linear-models-attempt-length-and-opti3

14. Variance inflation factors and partial R^2 value diagnostics for mixed effects models with all melodic features as predictors
Section 7:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#diagnostic-statistics-for-models-with-all-features-in-partial-r-squared-and-variance-inflation-factor-values

15. Information about melodic features used in this study
Section 2.2:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#description-and-distribution-of-melodic-features

16. An alternative by-participant visualization of our data
Section 4:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#average-by-participant-across-trials

17. Statistical models to support observations about changes in similarity as a function of melody section and attempt
Section 9:
https://sebsilas.github.io/musicassessr/articles/silas_and_mullensiefen_2023_online_supplement.html#statistical-models-to-support-changes-in-similarity-as-a-function-of-attempt-and-melody-section