Psychomusicology: Music, Mind, and Brain
2017, Vol. 27, No. 4, 297–311

0275-3987/17/$12.00 http://dx.doi.org/10.1037/pmu0000197

Interrater Agreement in Memory for Melody as a Measure of Listeners' Similarity in Music Perception

Steffen A. Herff and Roger T. Dean
Western Sydney University

Kirk N. Olsen
Western Sydney University and Macquarie University

Music is a cultural universal, yet the individual experience of music can strongly differ between listeners. Here, we investigate the similarity of listeners' response patterns in the context of memory for melody and argue that memory can serve as a proxy to perception. If music perception is similar across listeners, then this similarity should be reflected in similar memory response patterns toward a specific melody corpus. We used interrater agreement in melody recognition tasks as a window into how "similarly" listeners perceive music, and melodies in particular. Specifically, the data of 10 published melody recognition experiments were reanalyzed and findings indicate interrater agreement of up to $r = .70$. However, interrater agreement was strongly dependent on whether explicit recognition or indirect recognition in the form of perceived familiarity was measured, with explicit recognition showing higher agreement among listeners. Furthermore, the specific melody corpus and tuning system played a significant role, as did whether melodies consisted of pitch-only, rhythm-only, or both pitch and rhythm information. Results are interpreted in light of their practical implications for computational models of memory for melody. We argue that these findings provide strong evidence that mathematical models designed to predict human memory for melody should focus on musical features that combine rather than separate components of rhythm and melody, and with greater emphasis on musical features that are independent of the tuning system.

Keywords: music perception, memory for melody, predicting recognition, stimulus specificity, perceptual similarity

Supplemental materials: http://dx.doi.org/10.1037/pmu0000197.supp

Most theories of human memory assume that memory representations are based on perceptual experiences (Dennis & Humphreys, 2001; Hintzman, 1984, 1988; McClelland & Chappell, 1998; Paivio, 1969; Shiffrin & Steyvers, 1997; Tulving, 1972). For example, the first component in the Atkinson–Shiffrin memory model is a stimulus input that leads to the sensory register that detects and holds sensory information (i.e., a perception; Shiffrin & Atkinson, 1969). Another example, MINERVA 2, is based on a first experience or event (e.g., a perception; Hintzman, 1984). A recent regenerative multiple representations (RMR) conjecture delves further into the connection between memory and perception by describing a crucial link between prior experience, perception, and the subsequent formation of memories (Herff, Olsen, & Dean, 2017; Herff, Olsen, Dean, & Prince, 2017; Herff, Olsen, Prince, & Dean, 2017). Here, we aim to utilize the connection between perception and memory as stated by the RMR conjecture and suggest using human memory as a window into listeners' perception of music.

The RMR conjecture asserts that perception directly influences memory (Herff, Olsen, & Dean, 2017, also see Malmberg & Annis, 2012 for the idea that memory is guided by perception). According to the RMR conjecture, how we perceive the world around us depends on our prior experience. Furthermore, if we experience multiple perceptual experiences simultaneously, we also form multiple memory representations. For example, when we hear a melody, we are presented with multiple perceptual experiences such as rhythm and pitch sequence, although we are able to integrate these two streams into a coherent perception of melody. According to the RMR conjecture, multiple perceptual experiences of a stimulus are individually weighted to form multiple memory representations of a coherent whole. In the context of music for example, this means that if we hear a melody, we perceive rhythm and pitch sequence, and because we are familiar with the underlying rules, we are able to integrate these two streams into a coherent perception of a melody (Deutsch, 1986; Krumhansl, 1991; Schneider, 1997). All of these perceptual expe-

Steffen A. Herff and Roger T. Dean, The MARCS Institute for Brain, Behaviour and Development, Western Sydney University; Kirk N. Olsen, The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, and Department of Psychology, Macquarie University.

We thank Andrew Milne, Daniel Müllensiefen, and Lauren Fairley for constructive feedback on an earlier version. We thank Jon Prince, Barbara Tillmann, Mari Jones, and Kate Stevens for their willingness to share data.

The present work is based on Steffen A. Herff's dissertation posted in the Western Sydney University library. The present work reanalyzes data published by Herff, Olsen, and Dean (2017); Herff, Olsen, Dean, et al. (2017); and Herff, Olsen, Prince, et al. (2017).

Correspondence concerning this article should be addressed to Steffen A. Herff, The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Post: Locked Bag 1797, Penrith NSW 2751, Australia. E-mail: s.herff@westernsydney.edu.au

riences form then the basis of weighted multiple memory representations. Forming multiple memory representations provides the benefits that if they code at least partially redundant information, then they can regenerate each other, making memories more long-lasting and resilient. Indeed, in the context of music, recent long-term memory studies have shown that the ability to form multiple memory representations of melodies is linked to remarkable resilience to memory-disrupting phenomena, such as intervening items. In these studies, participants were asked to listen to melodies from either an unfamiliar (Herff, Olsen, Dean, et al., 2017) or a familiar (Herff, Olsen, & Dean, 2017) tuning system and resilience toward intervening item interference was measured. As listeners are familiar with underlying rules that govern melodies written in a familiar tuning system, they can perceive an integrated melody in addition to only pitch and rhythmic sequences and thereby form multiple representations of the melody, and thus more resilient memories (also see Herff, Olsen, Prince, et al., 2017 for further tests of the RMR conjecture). The conjecture suggests that differences in perceptual experience should translate to differences in memory representations. Consequently, similarities in memory representations are indicative of similarities in perceptual experience. This suggests that similarity in listeners' memory response patterns to a specific set of melodies may be seen as a window into similarities between listeners' perception of that specific set of melodies. As a result, the present reanalysis of 10 published melody recognition experiments assesses interrater agreement rather than absolute memory performance to investigate the question of how similarly listeners perceive music in the context of memory for melody. It is important to note that absolute recognition performance and interrater agreement are not necessarily correlated. Indeed, it is possible that multiple listeners produce a similar memory response pattern toward a set of stimuli and yet show low actual recognition performance. Specifically, here we addressed how interrater agreement varies between (a) melody corpora within the same tuning system, (b) between melody corpora with different tuning systems, (c) between explicit and indirect measurements of memory, and (d) between melodies that consist of pitch-only sequences, rhythm-only sequences, or both pitch and rhythm information combined. These manipulations are discussed in the following sections. As discussed later, we aim for the present results to provide practical information for computational models designed to predict memory for melody based on musical features.

Melody Corpora With the Same Tuning System

Even within a given tuning and tonal system, there are still substantial differences in musical materials. A simple example of this would be the large variety of different musical genres that are popular and commonly heard in the Western music tradition. It is possible that the degree of similarity in listeners' perception varies between different music styles, even if the tuning system is identical. Here, we address this by analyzing memory in response to two corpora of music that are both in the tuning system familiar to Western listeners and yet are distinctly different in their genre (see *Stimulus* section).

Melody Corpora With Different Tuning Systems

Testing melody corpora with different tuning systems is a compelling way of approximating the influence of musical enculturation on music perception and cognition (Stevens, 2012). This is because dissimilarities in tuning systems are readily apparent between musical cultures and most likely influence music perception (Stevens, Tardieu, Dunbar-Hall, Best, & Tillmann, 2013). Potentially, the degree of similarity in music perception between listeners will vary depending on whether the melodies use a set of pitch and interval rules that are familiar to listeners. Indeed, music traditions come with underlying rules and expectations (Deutsch, 1986; Krumhansl, 1991, p. 295). The degree of similarity in listeners' perception and memory of melodies may depend on not only the exact pitches used but also whether listeners are familiar with these underlying rules and expectations of melodies (Castellano, Bharucha, & Krumhansl, 1984; Morrison, Demorest, Aylward, Cramer, & Maravilla, 2003; Pearce & Wiggins, 2006).

Here, we analyze interrater agreement in participants' memory response patterns to melody corpora in three different tuning systems (see *Stimulus* section). Besides the tuning system familiar to Western listeners, we also use a novel tuning system that shares many of the underlying rules with the familiar tuning system; however, it uses a different pitch set. The third tuning system is a novel tuning system that uses unfamiliar rules as well as an unfamiliar pitch set.

Explicit and Indirect Measurements of Memory

Explicit information about the nature of a memory task can lead to different performances when compared with indirect memory tasks (Fleischman, Wilson, Gabrieli, Bienias, & Bennett, 2004; Gaudreau & Peretz, 1999; Halpern & O'Connor, 2000). This is also observed in musical memory tasks (Halpern & Bartlett, 2010). Here, we analyze interrater agreement in explicit *and* indirect memory tasks. In an explicit task, participants are instructed and therefore aware that their memory is being tested. In an indirect memory task, participants are not directly informed that they are participating in a memory task. As a result, listeners will vary in the degree to which they suspect that this task may contain a memory component. Considering that indirect memory tasks often entail high levels of response variability (Buchner & Wippich, 2000; LeBel & Paunonen, 2011; Ward, Berry, & Shanks, 2013b, but also see Ward, Berry, & Shanks, 2013a for further discussion) compared with explicit task instructions that function effectively in homogenizing listeners, we predict that explicit memory tasks show higher interrater agreements than indirect memory tasks.

Pitch-Only Sequences, Rhythm-Only Sequences, and Pitch–Rhythm Information Combined

Melodies usually consist of a pitch sequence combined with a rhythmic sequence. Generally, it appears to be easier to recognize a melody when hearing its pitch-only sequence rather than its rhythm-only sequence (Hébert & Peretz, 1997; Herff, Olsen, Prince, et al, 2017; White, 1960). However, recognition performance of a melody is best when the original combined pitch and rhythmic sequences is presented (Hébert & Peretz, 1997). Indeed, when comparing memory performance of isochronous melodies

with that for melodies with folksong rhythms, superior memory performance for melodies with nonisochronous rhythms show the importance of rhythmic information in memory (Dowling, Kwak, & Andrews, 1995).

A previous study tested and supported the RMR conjecture's hypothesis that rhythm-only and pitch-only sequences should not show less cumulative disruptive interference than combined versions of the same stimuli (Herff, Olsen, Prince, et al., 2017). Here, the present study reanalyzes the data of previous studies to compare interrater agreement in memory tasks for melodies that consist of combined pitch and rhythmic sequences with interrater agreement of memory tasks that test pitch-only sequences and rhythm-only sequences.[1] This comparison sheds light on whether the perception between listeners of pitch-only sequences, rhythm-only sequences, or combined pitch–rhythm melodies is more similar. The different contributions of pitch-only, rhythm-only, or both combined to the similarity in listener's perception may inform our general understanding of music perception and could inform computational models that predict melody recognition, as elaborated in the following sections.

Predictive Models of Memory

In the visual domain, "memorability" of a picture appears to be a stable property across human observers. This means that pictures that are more memorable for one person are also most likely more memorable for another person (Isola, Xiao, Torralba, & Olivia, 2011). In other words, observers show high interrater agreement in their memory response patterns toward a set of pictures. In the study by Isola et al. (2011), participants were split into two groups and the responses from one half were correlated with those from the other half ($r = .75$). The correlation coefficient can be squared to obtain the proportion of variance in the response pattern from one half of participants that can be explained with the response pattern from the other half (Cohen, 1988).

Recently, Flexer and Grill (2016) measured listeners' interrater agreement in a music similarity task in an attempt to validate computational models of music similarity (based solely on acoustic features). Similar to Isola et al. (2011), they also correlated the response patterns from one half of participants with the other half (see p. 244). They found average split-half correlations of $r = .40$. This means that ~16% of the variance in the pattern of average responses from one half of participants can be explained by the response pattern of the other half. The authors argue that this result has implications for computational models of music perception that predict average similarity responses to musical material. Specifically, as the interrater agreement decreases, so does the possible performance of a predictive model. In other words, the higher the interrater agreement between listeners for a category of perceptual response, the more precise a predictive model of that response could potentially be. This means that average interrater agreement may be useful to identify tasks or specific stimulus corpora that show promise for use in developing predictive models of perceptual responses. Similarly, the higher the interrater agreement is in a memory for melody task that uses a specific melody corpus, the more precise a predictive model of that corpus and that task could potentially be.

Predicting Memory for Melodies

Recognition of monosyllabic words can be predicted well (up to 45% of the variance in hit rates) using underlying features such as word frequency (Cortese, Khanna, & Hacker, 2010). In music, underlying melodic features can relate to popularity (Kopiez & Müllensiefen, 2011), and a recent study using a blocked indirect recognition design explained between 9.6% and 25.3% of the variance in the recognition responses using musical features (Müllensiefen & Halpern, 2014). Müllensiefen and Halpern (2014) developed computational models that predict average recognition performance of melodies based on musical features of the stimuli. Understanding which of these features carry predictive value can shed light on which features listeners may use to selectively base recognition judgments, both consciously or nonconsciously (Berenzweig, Logan, Ellis, & Whitman, 2004; Eerola, Jäärvinen, Louhivuori, & Toiviainen, 2001; Müllensiefen, 2009; Müllensiefen & Halpern, 2014; Pearce & Müllensiefen, 2017; Velardo, Vallati, & Jan, 2016). In the following sections, we will discuss the implications of the present work for computational models that, similar to Müllensiefen and Halpern (2014), aim to predict average memory performance for melodies based on musical features of the melodies.

One problem with models that use stimulus features to predict memory responses is that there are nonstimulus-related processes involved in memory. For example, variables such as decay over time, emotional associations, attention lapses, lack of motivation, fatigue, expertise, or repetition all influence memory for melody (Cuddy et al., 2012; Gardiner, Kaminska, Dixon, & Java, 1996; Herff & Czernochowski, 2017; McAuley, Stevens, & Humphreys, 2004; Samson, Dellacherie, & Platel, 2009). As a result, it is not always clear how much variance a model using stimulus features could potentially explain and how much variance is due to nonstimulus-related processes.

The present reanalysis aims to shed light on perceptual similarity between music listeners by investigating the average proportion of variance that a pattern of responses from one group of participants can explain the pattern of responses in another group. In general, the higher the proportion of explained variance between participant groups, the more promising the predictive model that is based on stimulus features. This is because the similarity in the average response pattern between participants toward a set of melodies can largely be attributed to stimulus features or some other constant feature of the experimental condition (e.g., this could be the environment, distractors, social circumstances experienced by participants, etc.), whereas dissimilarities in the response patterns can largely be attributed to nonstimulus-related interindividual differences.

Methods of Testing Memory

In their investigation of memory for melody, Müllensiefen and Halpern (2014) used a blocked recognition design. In a blocked design, participants first hear a large number of melo-

[1] Note that the present investigation is concerned with similarity in listeners' responses (for more details on recognition performance for pitch-only and rhythm-only sequences, see Herff, Olsen, Prince, et al., 2017).

dies in a learning phase and then hear new melodies mixed with old melodies in a test phase. Participants then have to decide which melodies have been previously presented. This paradigm is a useful tool for memory research, as it provides a clear distinction between encoding and retrieval. However, everyday music recognition does not usually provide a clear separation between encoding and retrieval. Rather, for every stimulus encountered, the same question of whether this stimulus has been heard before is assessed without specific focus on encoding or retrieval. A useful alternative to a blocked design is a continuous recognition paradigm (Shepard & Teghtsoonian, 1961). In a continuous recognition paradigm, participants are presented with one melody after another throughout the experiment and they judge if each melody has previously been presented in the experiment. Sometimes, a melody is presented that has indeed been presented before. This paradigm has the advantage that it does not provide participants with information about whether they have to focus on encoding or retrieval (Dowling, 1991). This is because a continuous recognition paradigm does not have distinct study and test phases.

A large body of research by Dowling and colleagues investigated short-term discrimination (up to ~30 s) of target melodies (transposed versions of melodies that have been presented before) from lures with same or different contours. Within the timeframe of up to ~30 s, they found memory decay (Dowling, Tillman, & Ayers, 2001) and interference (Dowling, 1991; Dowling et al., 1995; Dowling, Magner, & Tillmann, 2016) for novel melodies that were presented once before.

In a series of studies using long-term continuous recognition paradigms, Herff and colleagues (Herff, Olsen, & Dean, 2017; Herff, Olsen, Dean, et al., 2017; Herff, Olsen, Prince, et al., 2017) investigated context variables that might influence overall melody recognition and might blur the predictive power of musical features. Surprisingly, with up to 195 intervening melodies, these long-term studies demonstrated that the number of intervening melodies has effectively no disruptive impact on long-term melody recognition performance (Herff, Olsen, & Dean, 2017). Similar results in the context of temporal decay rather than interference have been obtained, showing that temporal delay of up to a week has minimal to no disruptive effect on melody recognition (Schellenberg & Habashi, 2015).

These findings provide additional motivation for the investigation of the predictive power of musical features in long-term melody recognition. This is because decay and interference are traditionally two of the major influences on memory (Criss, Malmberg, & Shiffrin, 2011; Deffenbacher, Carr, & Leu, 1981; Lew, Pashler, & Vul, 2016; Norman, 2013; Oberauer, Awh, & Sutterer, 2017; Oberauer & Lewandowsky, 2011; Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012). With the impact of these two variables shown to be minimal in memory for melodies, the proportion of variance explainable by musical features might be larger than using other stimuli where these two effects are important (e.g., words, Friedman, 1990; letter trigrams, Olson, 1969; prose, Tillmann & Dowling, 2007; pictures, Konkle, Brady, Alvarez, & Oliva, 2010; Nickerson, 1965; faces and numbers, Donaldson & Murdock, 1968; Sadeh, Ozubko, Winocur, & Moscovitch, 2014).

Experiments

Aim and Rationale

This article is an analysis of the combined data obtained by Herff and colleagues' investigations of memory for melody originally published (Herff, Olsen, & Dean, 2017; Herff, Olsen, Dean, et al., 2017; Herff, Olsen, Prince, et al., 2017). Here, we investigate whether the average proportion of variance that a pattern of responses from one group of participants can explain the pattern of responses in another group (in other words, interrater agreement; Flexer & Grill, 2016), comparable with intersubject correlation in functional MRI studies such as Pajula and Tohka (2016). We aim to shed light on the question "how similar is music perception between listeners?" We use memory as a proxy to address this question and assume that if multiple listeners' perception of music is similar, then response patterns (in this case, memory) to particular melody corpora will also be similar. We argue that the present findings will increase our understanding of music perception and inform future computational models that aim to use musical features to predict memory for melody.

Method

Participants. All participants were recruited from Western Sydney University (Experiments 1, 2, 3, 4, 5, and 6) or Murdoch University (Experiments 7, 8, 9, and 10), Australia. The first six experiments exclusively recruited participants with fewer than 2 years of formal musical training and who were not actively participating in any form of music. Experiments 7, 8, 9, and 10 recruited participants with mixed musical background, predominately consisting of nonmusicians. Undergraduate university students enrolled in psychology courses comprised the vast majority of participants.

Procedure. All studies analyzed here used the same basic continuous recognition paradigm. After each melody was presented, participants were required to make a response before the next melody started. In eight of the experiments, participants were informed that after each melody, they should judge whether they have heard this melody in this experiment before by pressing a button labeled "old" or whether this is the first time they have heard this melody by pressing a button labeled "new." In two experiments, instead of explicit memory task instructions, participants were only instructed to indicate perceived familiarity on a 100-point visual analogue scale. The increase in perceived familiarity between first and second presentation of the melodies provides an indirect approach to measure recognition performance. In six experiments, every melody was presented twice throughout the experiment (i.e., one repetition). In four experiments, every melody occurred three times (i.e., two repetitions). In these cases, results are reported separately for both melody repetitions.

Stimuli. A summary of the stimuli characteristics is provided in Table 1. More information about the stimuli is presented in Appendix—Stimuli. The stimuli of all experiments as well as musical feature analysis can also be found in the online supplemental material S1 Stimuli.zip.

Two experiments used melody corpora that resembled modern advertisement jingles, and another two corpora used European folk songs (Herff, Olsen, & Dean, 2017). We assume that our Austra-

Table 1
Summary of the Results

Experiment	*N*	Stimuli and memory task	*r*	*SD*	*d*	r^2
1	28	60, 12 s, familiar 12-TET, piano, sound like jingles, recognition	.631	.065	5.787	.409
2	20	55, 12 s, familiar 12-TET, piano, from Experiment 1, recognition	.561	.071	4.977	.3161
3	32	98, 10.86 s, familiar 12-TET, piano, European folk songs, recognition	.322	.067	3.665	.106
4	30	98, 10.86 s, familiar 12-TET, piano, same as Experiment 3, familiarity	.135[a]	.076	1.506	.023
5	37	50, 10.86 s, unfamiliar 88.08-CET, piano, based on Experiments 3 and 4, recognition	.340	.093	2.829	.134
6	27	50, 10.86 s, unfamiliar 88.08-CET, piano, same as Experiment 5, familiarity	−.056[a]	.104	.391	.015
7	105	37, 2.695 s, unfamiliar novel artificial tuning, pure tones, recognition	1st .710 2nd .594	.063 .081	5.372 4.630	1st .510 2nd .360
8	36	37, 2.695 s, unfamiliar novel artificial tuning, pure tones, same as Experiment 7, recognition	1st .613 2nd .539	.079 .087	4.770 4.102	1st .384 2nd .294
9	34	37, 2.695 s, unfamiliar novel artificial tuning, pure tones, pitch-only, based on Experiments 7 and 8, recognition	1st .146 2nd .180	.121 .120	1.024 1.192	1st .037 2nd .049
10	36	37, 2.695 s, unfamiliar novel artificial tuning, pure tones, rhythm-only, based on Experiments 7 and 8, recognition	1st .503 2nd .499	.094 .097	3.673 3.512	1st .258 2nd .256

Note. The table depicts from left to right, experiment index, number of participants, a short description of the stimuli (including number of melodies, melody duration, tuning system, timbre, addition notes), whether participants were instructed explicitly to make a recognition judgment or indirectly by reporting perceived familiarity, average split-half correlations, standard deviation of the average split-half correlations, Cohen's *d* as effect size (standardized distance to the correlations observed when melody names are shuffles in one half of the participants), and the average of all squared split-half correlations. 12-TET refers to the 12-tone-equal-temperament tuning system, the tuning system most dominant in Western cultures. 88-CET refers to a novel, artificial tuning system that uses an 88 cent step equal temperament, rather than the 100 cent in 12-TET. The novel tuning system of Experiments 7 to 10 is detailed in the text. Note that Experiments 7–10 repeated melodies twice throughout the experiments. In these cases, results for both repetitions are reported separately.

[a] Note that average interrater agreement only changes minimally when the 100-point familiarity rating scale is binned into 10 s. In Experiment 4, $r = .138$, $SD = .074$, when binned into 10 s. In Experiment 6, $r = -.071$, $SD = .106$, when binned into 10 s.

lian listeners are generally more familiar with the style of the corpus that resembles jingles compared with the European folk song corpus (Herff, Olsen, & Dean, 2017). This means that a total of four experiments used stimuli in 12-tone-equal-temperament (12-TET), the tuning system familiar to Western listeners. The other six experiments used melodies in unfamiliar tuning systems. Some of the melodies in an unfamiliar tuning system were based on the melodies in a familiar tuning system. These experiments realized the melodies in a new 88-cent-equal-temperament (CET) tuning system. The 88-CET tuning system is the equally tempered tuning system most dissimilar to the familiar Western tonal system within all 40- to 100-CET tuning systems (Herff, Olsen, Dean, et al., 2017) based on the tonal affinity model of Milne et al. (Milne, 2013; Milne & Holland, 2016; Milne, Laney, & Sharp, 2015, 2016; Milne, Sethares, Laney, & Sharp, 2011). This means that adjacent pitches in the tuning system are 88 cents apart, rather than the 100 in the Western tonal tuning system. As a result, this tuning system (Experiments 5 and 6) uses a different pitch set; however, the melodies have the same contour and rhythm than familiar melodies (Experiments 3–5), only the pitch of each note is adjusted to the new tuning system.

The melodies in Experiments 7 to 10 did not conform to Western music tradition in rhythm and tonality (Herff, Olsen, Dean, et al., 2017; Herff, Olsen, Prince, et al., 2017). The melodies used pitch heights of 480, 520, 560, 605, and 665 Hz and note durations of 60, 110, 550, and 920 ms, with a 100-ms silent gap between notes.[2] Some experiments presented this set of stimuli either in pitch-only (isochronous rhythm with pitch variations in Experiment 9), rhythm-only (same pitch but nonisochronous rhythm in Experiment 10), or combined versions (Experiments 7 and 8) with both pitch and rhythm information. This allows the present analysis to compare the contribution of rhythm and pitch sequences to the recognition and perceived familiarity response similarity between participants. As stated before, the present article is concerned with interrater agreement rather than memory accuracy. Memory accuracy results for the present data set can be found in Herff, Olsen, and Dean (2017) for Experiments 1 to 4; in Herff, Olsen, Dean, et al. (2017) for Experiments 5 to 7; and in Herff, Olsen, Prince, et al. (2017) for Experiments 8 to 10.

Statistical approach. We assess interrater agreement (or similarity) in recognition by samples of participants toward sets of melodies. To this end, in each experiment, the average recognition

[2] In pitch perception and a standard/comparison task with a silent retention interval, a Weber fraction of .04 led to discrimination performance above 90%. The pitch Weber fractions used in the present stimuli were between .08 and .10, therefore clearly discriminable. In duration perception, Weber fractions of .30 led to discrimination performance above 98%. The duration fractions used in this study ranged from .67 to 1.27, therefore also clearly discriminable. The melodies were constructed in various artificial grammars, unfamiliar to listeners, which were of importance for the later investigation of statistical learning that will be reported elsewhere.

performance for each melody of half the participant sample was correlated with the average recognition performance of the other half toward the same melodies. Specifically, the entire participant sample was split into two halves and the average hit rate for each melody was calculated. These average hit rates for each melody were then correlated with the average hit rates for the same melodies by the other half of the sample. To increase estimation precision of the average split-half correlation, we repeated this process 1,000 times with random splits and average split-half correlations. This means that each resampling uses the data of all participants within one experiment; however, each participant's data become newly randomly allocated to one of the two split-halves during each of the 1,000 repetitions of the procedure. (This procedure is identical to Isola et al.'s, 2011, analysis of interrater agreement and similar to Flexer & Grill's, 2016, analysis of interrater agreement, with the addition of resampling to increase estimate precision and averaging across participants.) The p values and standard deviations are reported for each experiment. The p values and Cohen's d (Cohen, 1988) were obtained by comparing the vector of actual split-half correlations with a vector of split-half correlations in which melody names were shuffled for one of the halves. Note that owing to the large number of (split-half) observations (1,000), most effects will show significance. Therefore, we use correlation values as interpretable effect sizes and report Cohen's d as a measure of the effect size in standard (Cohen, 1988).

Furthermore, the proportion of variance that the response pattern of one group of participants predicts from the response pattern of another can inform predictive models of memory that use stimulus features as predictors. Specifically, the higher the proportion of explainable variance, the more promising the predictive model will be (Flexer & Grill, 2016). Here, we use average split-half r^2 as an indicator of the average proportion of variance that the response pattern of one group of participants predicts in the response pattern of another (Cohen, 1988; Flexer & Grill, 2016). Average split-half r^2 values were calculated by squaring each of the 1,000 split-half correlations and averaging them. All experiments yield significant average split-half correlations of melody-specific recognition with 1,000 splits (all $p < .0001$) and effects can be directly interpreted from the figures. Figures will depict standard deviations instead of standard errors and confidence intervals. This is because with the large number of split-half correlations, standard errors become too small to be informative in visual representations of the data.

Results

A summary of the results can be found in Table 1. The table reports average split-half correlations for all 10 experiments. Note that Experiments 5 and 6 used the same melodies previously used in Experiments 3 and 4, but retuned to an unfamiliar tuning system. Experiments 4 and 6 used the identical stimuli compared with Experiments 3 and 5, respectively, but in an indirect task where participants were instructed to report perceived familiarity, rather than explicit recognition. Figure 1 shows the split-half correlation distributions for each experiment. In general, the wider a distribution is, the larger the differences between the 1,000 split-halves, whereas a narrow distribution represents more similar split-half correlations in all 1,000 split-halves (the spread is also represented in the *SD* column of Table 1). Similarly, the farther left a distribution is, the lower the split-half correlations, whereas distributions to the far right suggest high split-half correlations (the mean of the distributions is also represented in the r column of Table 1).

Task Instructions and Tuning Systems

Figure 2 depicts means and standard deviations of the average split-half correlations and visualizes the following two main findings: (a) the indirect familiarity tasks show lower split-half correlations compared with the recognition tasks, and (b) the unfamiliar tuning system showed slightly higher split-half correlations compared with the familiar tuning system. Furthermore, Figure 2 depicts the interaction between *Tuning system* and *Task instructions*, showing that the familiar tuning system produces higher interrater agreement in the indirect perceived familiarity task compared with the familiar tuning system, but this is not observed when an explicit recognition task is used. Note that Figure 2 only shows the data of Experiments 3, 4, 5, and 6. This is because these four experiments use the same timbre (piano), the same underlying melody corpus (Experiments 3 and 4 tuned to 12-TED and Experiments 5 and 6 tuned to 88-CET) and the same recognition tasks (Experiments 3 and 5 use recognition judgments and Experiments 4 and 6 use perceived familiarity judgments), enabling direct investigation of the effect of task instruction and tuning system on interrater agreement in the form of average split-half correlations.

Repetition, Pitch-Only, Rhythm-Only, and Combined Sequences

Figure 3 shows average split-half correlations in regards to the different types of stimuli after the first and second repetition. Overall, rhythm-only sequences show higher average split-half correlations ($M_r = .501$, $SD_r = .096$) than the pitch-only sequences ($M_r = .163$, $SD_r = .122$). The combined melodies show more explainable variance ($M_r = .614$, $SD_r = .099$) than did their underlying pitch and rhythmic sequences separately.

Discussion

The present study analyzed the memory response patterns of listeners to various melody corpora from 10 different experiments. We calculated interrater agreement of memory recognition between participants. Assuming that listeners' similarity in memory response patterns is primarily driven by perceptual similarities between listeners, we investigate the extent to which listeners' music perception converges. We observed some striking similarities between listeners with interrater agreement of up to $r = .71$. The highest average interrater agreement we observed is similar to those in other domains (e.g., monosyllabic recognition, $r = .70$, in Cortese et al., 2010, and picture recognition, $r = .75$, in Isola et al., 2011). However, average interrater agreement significantly varied based on memory task instructions, tuning system, and nature of the stimuli (pitch-only, rhythm-only, both combined).

In light of these main findings, the degree of listeners' similarity in music perception as well as the implications for predictive models of melody recognition will now be discussed in the context of the influence of melody corpora, the influence of memory task instructions, the influence of the tuning system, and the relative

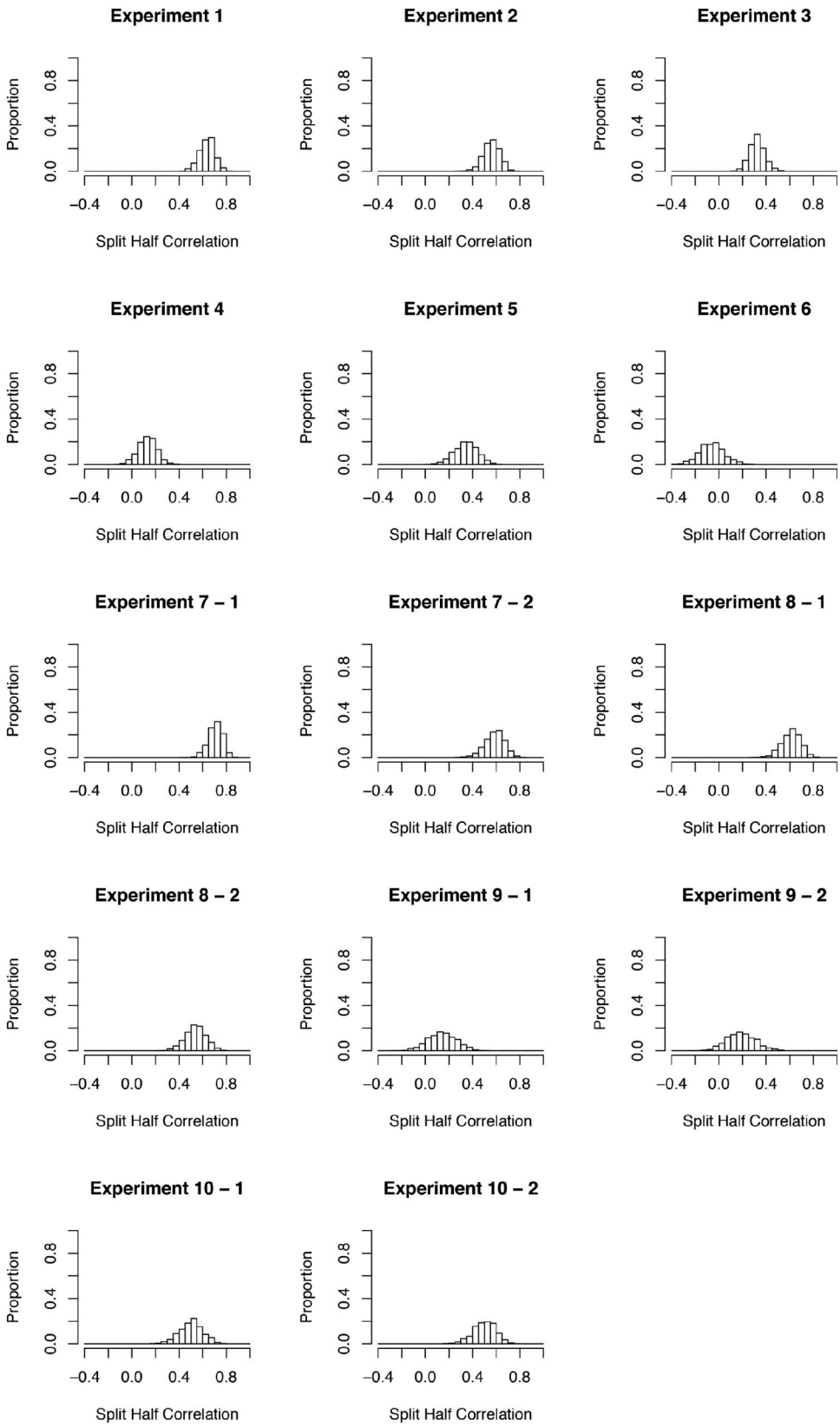

Figure 1. Distributions of split-half correlations in each experiment. For Experiments 7 to 10, the dashed numbers "1" and "2" shown after the title of each experiment index to the first and second repetition of the melodies.

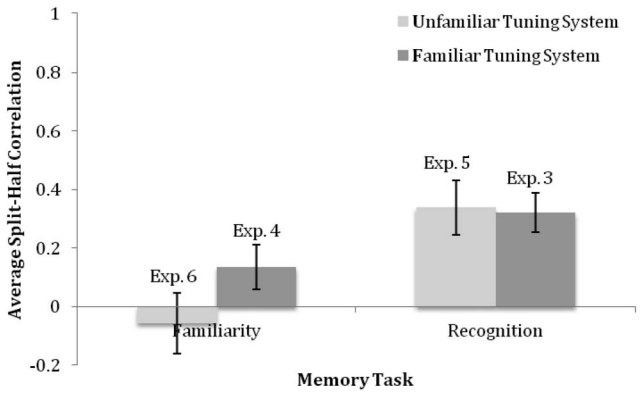

Figure 2. Data of Experiments 3, 4, 5, and 6. Average split-half correlations for the indirect perceived familiarity tasks, the explicit recognition task, and both unfamiliar and familiar tuning systems. Error bars show standard deviation because standard errors (and confidence intervals) are too small to depict due to the large number of split-half correlations.

contribution of rhythm and pitch sequence to stability of memorability of melodies.

The Influence of Melody Corpora

Within the Western-tonal experiments analyzed in the present study, there was a large similarity between listeners' recognition response patterns. However, the degree of agreement varied between corpora, even though these corpora used melodies in the same tuning system. A new corpus resembling modern advertisement or movie themes (Experiments 1 and 2) showed higher interrater agreement than a corpus of European folk songs (Experiment 3). This suggests that similarity in listeners' perception of music changes as a function of the precise auditory material. Depending on the melody corpus, listeners are therefore likely to be very similar in the way they perceive the music. More precisely, it suggests that similarity in music perception decreases and memory response patterns diverge as the musical style of a melody corpus becomes less familiar. This may be because music that resembles advertisements and movie themes are likely to be more familiar or at least more easily remembered for Australian listeners than European folk songs (see Herff, Olsen, & Dean, 2017, for precise recognition performance). However, as the melody corpus becomes even more unfamiliar, memory response patterns seem to converge again (see Experiment 7). Taken together, this could suggest that similar degrees of familiarity with a melody corpus between the listeners lead to higher interrater agreement, as listeners may have been similarly familiar with style of melodies used in Experiments 1 and 2 and similarly unfamiliar with the artificial grammar that is the bases of the melodies in Experiment 7, leaving large differences in familiarity between participants on Experiment 3's European folk melodies, which may have been familiar to some but not others. This hypothesis could be specifically addressed in the future, by testing interrater agreement as a function of group coherence in terms of familiarity with the melody corpus.

The similarity between participants' recognition judgments also informs computational models that aim to predict average memory responses using musical features as predictors (similar to Flexer & Grill, 2016; Müllensiefen & Halpern, 2014). By squaring the average split-half correlation coefficients, we calculated the proportion of variance that the response pattern from one group of participants explains of the response pattern from another group (Cohen, 1988; Flexer & Grill, 2016). In Experiment 1, the present results show that an average proportion of up to ~40% of the variance can be explained by the response pattern of another group of participants. This suggests that the endeavor to develop musical feature models that predict recognition patterns to a substantial degree can be a feasible one, as a large proportion of the variance seems to be stimulus-driven. This assertion is based on the rationale that a large proportion of the unexplainable variance between two groups of participants who perform the same task is based on interindividual differences, whereas a large proportion of the variance that can be explained is based on stimulus features. In a practical context, we argue that the higher the proportion of explained variance between two participant groups, the more promising the predictive model will be when memory is predicted by stimulus features within the music.

The wide range of variance that can be explained (10%–40% in Experiments 1, 2, and 3) for responses within the Western melody corpora also shows that the potential usefulness of predictive feature models depends on the exact melody corpus. As the corpora that approximated pop music or advertisement melodies elicited greater interrater agreement in memory responses when compared with traditional European Folk melodies, the present results point toward future questions. Specifically, future research could address the question of the influence of genre familiarity on similarity in listeners' perception. Future studies could try to identify which musical features in particular are responsible for the differences in interrater agreement of different melody corpora. However, such an endeavor should be wary of differences in sample expertise, as a group with high level of musical expertise may have a different response profile. This is because musicians may have similar degrees of musical expertise but entirely different specializations that influence their perception. In general, it is

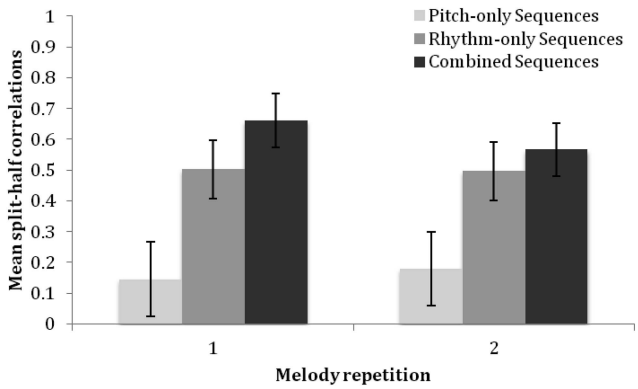

Figure 3. Data of Experiments 7, 8, 9, and 10. Average split-half correlations for both melody repetitions and all three kinds of stimuli, the original melodies in an unfamiliar tuning system that consisted of a combined pitch and rhythmic sequence, as well as underlying pitch and rhythmic sequences tested separately in a new sample. Error bars show standard deviation because standard errors (and confidence intervals) are too small to depict due to the large number of split-half correlations.

to be expected that the more homogenous a participant sample is, the more similar their responses will be. As a result, this would increase the proportion of explainable variance but simultaneously make predictive models less generalizable. In terms of the generalizability of a predictive model, it is also important to consider the precise memory task utilized.

The Influence of Memory Task

The present reanalysis used data from multiple experiments that deployed identical melody corpora but different memory tasks. All experiments analyzed here used a continuous recognition paradigm, but in some experiments, participants were instructed that some melodies would be repeated throughout the experiment, and therefore, they were required to report which melodies had been presented before (explicit memory task). In other experiments, participants were instructed to just rate their perceived familiarity toward the melodies (indirect memory task). Even though sometimes the same overall memory phenomena can be demonstrated in both the indirect and explicit memory tasks (Herff, Olsen, & Dean, 2017; Herff, Olsen, Dean, et al., 2017), it is clear from the present findings that the different tasks elicit significantly different interrater agreement in response to the same melody corpora. Memory responses were more similar between participants in the recognition task than the perceived familiarity tasks, even though the same melodies were used. This finding is not surprising considering that indirect paradigms may introduce additional participant-wise variation that comes with the uncertainty of task instructions. Nevertheless, this is worth noting for future experiments and analyses, as this is also in line with previous findings (Blaxton, 1989; Gopie, Craik, & Hasher, 2011; Müllensiefen & Halpern, 2014; Richardson-Klavehn, Gardiner, & Java, 1996).

The substantially smaller interrater agreement reported here between groups of participants in the indirect memory task also indicates less potentially explainable variance for musical feature models designed to predict perceived familiarity. This suggests that modeling recognition instead of perceived familiarity may be a more effective endeavor when aiming to explain large proportions of variance. These results also suggest that future investigations utilizing indirect measures of memory (such as increases in perceived familiarity between melody occurrences) to build predictive models of melody recognition might not be as fruitful as they intuitively seem. Interestingly, the effect size from comparing differences in the average split-half correlations between first and second melody repetitions (Experiments 7–10) was negligibly small ($\eta_p^2 = .011$). This means that the degree of interrater agreement (not the performance) is hardly affected by the additional melody repetition, suggesting that a small number of additional melody repetitions do not increase similarity between listeners' perception of a piece. More empirical attention, however, will be useful to evaluate models that utilize tuning system-independent musical features as the basis of their memory response predictions.

The Influence of Tuning System

The present investigation comprises experiments that presented melodies in the tuning system familiar to participants and tuning systems that were unfamiliar to participants. Importantly, Experiment 5 used stimuli directly based on Experiment 3, only detuned into an unfamiliar tuning system. As a result, many musical features such as rhythm and pitch contour in the stimulus set were identical between these experiments. Interestingly, interrater agreement was comparable between these two corpora. This result suggests that the degree of similarity in listeners' music perception does not change dramatically with the tuning system per se but, rather, with musical features that are unaffected by the tuning system.[3] Consequently, it may be that a large proportion of the explainable variance in memory for melodies might be due to musical features that are independent of the tuning system (such as melody contour). This observation has direct implications for present and future models aiming to predict melody recognition (Müllensiefen & Halpern, 2014), as musical features that evoke similar perceptual responses in listeners may also carry large proportions of predictive power. However, it is important to note that only the average split-half correlations between the two corpora were comparable in the direct melody recognition task. In the indirect perceived familiarity task, the familiar tuning (Experiment 4) produced higher average split-half correlations compared with the unfamiliar tuning (Experiment 6). Further support that underlying musical features might predict melody recognition performance, even outside the domain of familiar tuning systems, can be observed from the data of Experiments 7 and 8. These experiments used a different unfamiliar tuning system and show proportions of explainable variance of ~37%. Taken together, the present findings suggest that the intrinsic predictive power of stimuli in melody recognition tasks might not derive from familiarity with the underlying tuning system but instead from musical features that operate independent of the tuning system.

The Influence of Rhythm and Pitch Sequence

The unique composition of stimuli used in Experiments 7, 8, 9, and 10 allow some additional conclusions. Experiments 9 and 10 used the same stimuli as Experiments 7 and 8. However, Experiment 9 provided participants solely with pitch-only versions of the stimuli and Experiment 10 used rhythm-only versions of the original stimuli in Experiments 7 and 8. This means that Experiments 7 and 8 provide a baseline in which to compare responses to combined rhythm and pitch sequences with pitch (Experiment 9) or rhythm (Experiment 10) sequences separately.

Interestingly, the degree of similarity in listeners' response pattern was higher in the rhythm-only sequences compared with the pitch-only sequences. Pitch-only sequences are often reported to be more memorable than rhythm-only sequences (Hébert & Peretz, 1997; White, 1960). The present analysis does not challenge these findings but, rather, suggests that participants show higher interrater agreement in their recognition judgments toward pure rhythmic sequences compared with pure pitch sequences.

Furthermore, the proportion of explainable variance was much higher in the combined sequences (~38%) compared with the pitch-only sequence (~4%), the rhythm-only sequence (~26%),

[3] We found further support for this interpretation in a post hoc analysis. We used the same average correlation procedure to correlate the response patterns in Experiment 3 (Western tonal) with Experiment 5 (melodies based on Experiment 3 but played in an unfamiliar tuning system). We found similar average split-half correlations between Experiments 3 and 5 ($r = .234$, $SD = .118$) relative to those within Experiment 3 ($r = .32$, $SD = .067$) and within Experiment 5 ($r = .34$, $SD = .09$).

and individual sequences as well as their sum (~30%). This suggests there are remarkable interactions between rhythm and pitch for listeners' similarity in perception. A candidate mechanism for these interactions could be the bilateral guidance of attention and phrase perception in rhythm and pitch. For example, rhythms can guide attention to specific parts of a melody and vice versa (Jones & Boltz, 1989; Jusczyk & Krumhansl, 1993; Palmer & Krumhansl, 1987; Schmuckler & Boltz, 1994). More importantly, such interactions appear to have similar effects between perceivers. Considering that these apparent interactions appear to account for a large proportion of the explainable variance of melodies, future attempts to predict memory for melody should investigate musical features that reflect interactions of rhythm and pitch in addition to features that reflect the individual contribution of rhythm and pitch (see Hébert & Peretz, 1997, for further information on the contribution of pitch and time to melody recognition; see Prince, 2014, for more information on the contribution of pitch and time to melodic similarity). This finding appears intuitive in the context of the RMR conjecture, which asserts that multiple perceptual experiences lead to multiple memory representations (Herff, Olsen, & Dean, 2017; Herff, Olsen, Dean, et al., 2017; Herff, Olsen, Prince, et al., 2017), and prior knowledge informs how and if these representations are integrated into a coherent whole representation. In other words, all listeners have a perceptual experience of the rhythm, and of the pitch sequence. In addition, if they have prior knowledge about how to integrate time and pitch, they will also have a perception and, therefore, memory representation of the integrated melody. Given that our participants all derived from a similar cultural background (Australian), it can be assumed that the way in which they integrate additional information into a coherent new melody representation would be similar. In turn, this would explain why multiple representations increase perceptual similarity between observers.

Conclusion

The present study used interrater agreement in the memory response patterns of listeners as a window into similarity of music perception. The results can inform predictive models of melody recognition that use musical features as predictors. Overall, results suggest that future models that aim to predict melody recognition based on musical features should carefully consider the precise task instructions given to the participants and focus on musical features that describe the interaction between rhythm and melody, as well as musical features that are tuning-system-independent. A model that can predict melody recognition beyond those variables using the commonalities in the predictive power of musical features would provide a strong framework for future research endeavors that aim to predict not only memory for melody but also memory for complete musical pieces.

References

Berenzweig, A., Logan, B., Ellis, D. P. W., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal, 28,* 63–76. http://dx.doi.org/10.1162/014892604323112257

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 657–668. http://dx.doi.org/10.1037/0278-7393.15.4.657

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40,* 227–259. http://dx.doi.org/10.1006/cogp.1999.0731

Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of north India. *Journal of Experimental Psychology: General, 113,* 394–412. http://dx.doi.org/10.1037/0096-3445.113.3.394

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18,* 595–609.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316–326. http://dx.doi.org/10.1016/j.jml.2011.02.003

Cuddy, L. L., Duffin, J. M., Gill, S. S., Brown, C. L., Sikka, R., & Vanstone, A. D. (2012). Memory for melodies and lyrics in Alzheimer's disease. *Music Perception, 29,* 479–491. http://dx.doi.org/10.1525/mp.2012.29.5.479

Deffenbacher, K. A., Carr, T. H., & Leu, J. R. (1981). Memory for words, pictures, and faces – retroactive interference, forgetting, and reminiscence. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 299–305. http://dx.doi.org/10.1037/0278-7393.7.4.299

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. http://dx.doi.org/10.1037/0033-295X.108.2.452

Deutsch, D. (1986). Auditory pattern recognition. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. II. Cognitive processes and performance* (pp. 32.31–32.49). New York, NY: Wiley.

Donaldson, W., & Murdock, B. B., Jr. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology, 76,* 325–330. http://dx.doi.org/10.1037/h0025510

Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception and Psychophysics, 50,* 305–313. http://dx.doi.org/10.3758/BF03212222

Dowling, W. J., Kwak, S., & Andrews, M. W. (1995). The time course of recognition of novel melodies. *Perception and Psychophysics, 57,* 136–149. http://dx.doi.org/10.3758/BF03206500

Dowling, W. J., Magner, H., & Tillmann, B. (2016). Memory improvement with wide-awake listeners and with nonclassical guitar music. *Psychomusicology: Music, Mind, and Brain, 26,* 26–34. http://dx.doi.org/10.1037/pmu0000106

Dowling, W. J., Tillman, B., & Ayers, D. F. (2001). Memory and the experience of hearing music. *Music Perception, 19,* 249–276. http://dx.doi.org/10.1525/mp.2001.19.2.249

Eerola, T., Jäärvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception, 18,* 275–296. http://dx.doi.org/10.1525/mp.2001.18.3.275

Fleischman, D. A., Wilson, R. S., Gabrieli, J. D. E., Bienias, J. L., & Bennett, D. A. (2004). A longitudinal study of implicit and explicit memory in old persons. *Psychology and Aging, 19,* 617–625. http://dx.doi.org/10.1037/0882-7974.19.4.617

Flexer, A., & Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research, 45,* 239–251. http://dx.doi.org/10.1080/09298215.2016.1200631

Friedman, D. (1990). ERPs during continuous recognition memory for words. *Biological Psychology, 30,* 61–87. http://dx.doi.org/10.1016/0301-0511(90)90091-A

Gardiner, J. M., Kaminska, Z., Dixon, M., & Java, R. I. (1996). Repetition of previously novel melodies sometimes increases both remember and know responses in recognition memory. *Psychonomic Bulletin and Review, 3,* 366–371. http://dx.doi.org/10.3758/BF03210762

Gaudreau, D., & Peretz, I. (1999). Implicit and explicit memory for music in old and young adults. *Brain and Cognition, 40,* 126–129.

Gopie, N., Craik, F. I., & Hasher, L. (2011). A double dissociation of implicit and explicit memory in younger and older adults. *Psychological Science, 22,* 634–640. http://dx.doi.org/10.1177/0956797611403321

Halpern, A. R., & Bartlett, J. C. (2010). Memory for melodies. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception* (Vol. 36, pp. 233–258). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-6114-3_8

Halpern, A. R., & O'Connor, M. G. (2000). Implicit memory for music in Alzheimer's disease. *Neuropsychology, 14,* 391–397. http://dx.doi.org/10.1037/0894-4105.14.3.391

Hébert, S., & Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory and Cognition, 25,* 518–533. http://dx.doi.org/10.3758/BF03201127

Herff, S. A., & Czernochowski, D. (2017). The role of divided attention and expertise in melody recognition. *Musicae Scientiae*. Advance online publication. http://dx.doi.org/10.1177/1029864917731126

Herff, S. A., Olsen, K. N., & Dean, R. T. (2017). Resilient memories for melodies: The number of intervening melodies does not influence novel melody recognition. *The Quarterly Journal of Experimental Psychology*. Advance online publication. http://dx.doi.org/10.1080/17470218.2017.1318932

Herff, S. A., Olsen, K. N., Dean, R. T., & Prince, J. (2017). Memory for melodies in unfamiliar tuning systems: Investigating effects of recency and number of intervening items. *The Quarterly Journal of Experimental Psychology*. Advance online publication. http://dx.doi.org/10.1080/17470218.2017.1333519

Herff, S. A., Olsen, K. N., Prince, J., & Dean, R. T. (2017). Interference in memory for pitch-only and rhythm-only sequences. *Musicae Scientiae*. Advance online publication. http://dx.doi.org/10.1177/1029864917695654

Hintzman, D. L. (1984). Minerva-2 - a simulation-model of human-memory. *Behavior Research Methods, Instruments and Computers, 16,* 96–101. http://dx.doi.org/10.3758/BF03202365

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. http://dx.doi.org/10.1037/0033-295X.95.4.528

Isola, P., Xiao, J., Torralba, A., & Olivia, A. (2011). What makes an image memorable? In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 145–152). Providence, RI: IEEE.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96,* 459–491. http://dx.doi.org/10.1037/0033-295X.96.3.459

Jusczyk, P. W., & Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 627–640. http://dx.doi.org/10.1037/0096-1523.19.3.627

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139,* 558–578. http://dx.doi.org/10.1037/a0019165

Kopiez, R., & Müllensiefen, D. (2011). Auf der suche nach den 'popularitätsfactoren' in den song-melodien des Beatles-albums revolvers [in search of popularity factors in the tunes from the album Revolver of the Beatles]. In S. Meine & N. Noeske (Eds.), *Musi and popularität: Aspecte zu einer kulturgeschichte zwischen 1500 und heute* (pp. 207–225). Münster, Germany: Waxmann.

Krumhansl, C. L. (1991). Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology, 42,* 277–303. http://dx.doi.org/10.1146/annurev.ps.42.020191.001425

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37,* 570–583. http://dx.doi.org/10.1177/0146167211400619

Lew, T. F., Pashler, H. E., & Vul, E. (2016). Fragile associations coexist with robust memories for precise details in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 379–393. http://dx.doi.org/10.1037/xlm0000178

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141,* 233–259. http://dx.doi.org/10.1037/a0025277

McAuley, J. D., Stevens, C., & Humphreys, M. S. (2004). Play it again: Did this melody occur more frequently or was it heard more recently? The role of stimulus familiarity in episodic recognition of music. *Acta Psychologica, 116,* 93–108. http://dx.doi.org/10.1016/j.actpsy.2004.02.001

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760. http://dx.doi.org/10.1037/0033-295X.105.4.734-760

Milne, A. J. (2013). *A computational model of the cognition of tonality* (Doctoral dissertation), The Open University, Milton Keynes, United Kingdom.

Milne, A. J., & Holland, S. (2016). Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music: Mathematical and Computational Approaches to Music Theory, Analysis, Composition and Performance, 10,* 59–85.

Milne, A. J., Laney, R., & Sharp, D. B. (2015). A spectral pitch class model of the probe tone data and scalic tonality. *Music Perception, 32,* 364–393. http://dx.doi.org/10.1525/mp.2015.32.4.364

Milne, A. J., Laney, R., & Sharp, D. B. (2016). Testing a spectral model of tonal affinity with microtonal melodies and inharmonic spectra. *Musicae Scientiae, 20,* 465–494. http://dx.doi.org/10.1177/1029864915622682

Milne, A. J., Sethares, W. A., Laney, R., & Sharp, D. B. (2011). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music, 5,* 1–20. http://dx.doi.org/10.1080/17459737.2011.573678

Morrison, S. J., Demorest, S. M., Aylward, E. H., Cramer, S. C., & Maravilla, K. R. (2003). FMRI investigation of cross-cultural music comprehension. *NeuroImage, 20,* 378–384. http://dx.doi.org/10.1016/S1053-8119(03)00300-8

Müllensiefen, D. (2009). *Fantastic: Feature analysis technology accessing statistics (in a corpus)*. Technical report v1. London, United Kingdom: Goldsmiths University of London.

Müllensiefen, D., & Halpern, A. R. (2014). The role of features and context in recognition of novel melodies. *Music Perception: An Interdisciplinary Journal, 31,* 418–435. http://dx.doi.org/10.1525/mp.2014.31.5.418

Nickerson, R. S. (1965). Short-term-memory for complex meaningful visual configurations – A demonstration of capacity. *Canadian Journal of Psychology, 19,* 155–160. http://dx.doi.org/10.1037/h0082899

Norman, D. A. (2013). *Models of human memory*. New York, NY: Elsevier.

Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: Proactive facilitation but no proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 1–22. http://dx.doi.org/10.1037/xlm0000302

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: A computational implementation of the time-based resource-sharing theory. *Psychonomic Bulletin and Review, 18,* 10–45. http://dx.doi.org/10.3758/s13423-010-0020-6

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin and Review, 19,* 779–819. http://dx.doi.org/10.3758/s13423-012-0272-4

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

Olson, G. M. (1969). Learning and retention in a continuous recognition task. *Journal of Experimental Psychology, 81,* 381–384. http://dx.doi.org/10.1037/h0027756

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review, 76,* 241–263. http://dx.doi.org/10.1037/h0027272

Pajula, J., & Tohka, J. (2016). How many is enough? Effect of sample size in inter-subject correlation analysis of fMRI. *Computational Intelligence and Neuroscience, 2016,* article id: 2094601. http://dx.doi.org/10.1155/2016/2094601

Palmer, C., & Krumhansl, C. L. (1987). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception and Psychophysics, 41,* 505–518. http://dx.doi.org/10.3758/BF03210485

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

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception, 23,* 377–405. http://dx.doi.org/10.1525/mp.2006.23.5.377

Prince, J. B. (2014). Contributions of pitch contour, tonality, rhythm, and meter to melodic similarity. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 2319–2337. http://dx.doi.org/10.1037/a0038010

Richardson-Klavehn, A., Gardiner, J. M., & Java, R. I. (1996). Memory: Task dissociations, process dissociations and dissociations of consciousness. In G. D. M. Underwood (Ed.), *Implicit cognition* (pp. 85–158). New York, NY: Oxford University Press.

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2014). How we forget may depend on how we remember. *Trends in Cognitive Sciences, 18,* 26–36. http://dx.doi.org/10.1016/j.tics.2013.10.008

Samson, S., Dellacherie, D., & Platel, H. (2009). Emotional power of music in patients with memory disorders: Clinical implications of cognitive neuroscience. *Annals of the New York Academy of Sciences, 1169,* 245–255. http://dx.doi.org/10.1111/j.1749-6632.2009.04555.x

Schellenberg, E. G., & Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Memory and Cognition, 43,* 1021–1031. http://dx.doi.org/10.3758/s13421-015-0519-1

Schmuckler, M. A., & Boltz, M. G. (1994). Harmonic and rhythmic influences on musical expectancy. *Perception and Psychophysics, 56,* 313–325. http://dx.doi.org/10.3758/BF03209765

Schneider, A. (1997). "Verschmelzung", tonal fusion, and consonance: Carl Stumpf revisited. In M. Leman (Ed.), *Music, gestalt, and computing: Studies in cognitive and systematic musicology* (pp. 117–143). London, United Kingdom: Springer. http://dx.doi.org/10.1007/BFb0034111

Shepard, R. N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology, 62,* 302–309. http://dx.doi.org/10.1037/h0048606

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review, 76,* 179–193. http://dx.doi.org/10.1037/h0027277

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin and Review, 4,* 145–166. http://dx.doi.org/10.3758/BF03209391

Stevens, C. J. (2012). Music perception and cognition: A review of recent cross-cultural research. *Topics in Cognitive Science, 4,* 653–667. http://dx.doi.org/10.1111/j.1756-8765.2012.01215.x

Stevens, C. J., Tardieu, J., Dunbar-Hall, P., Best, C. T., & Tillmann, B. (2013). Expectations in culturally unfamiliar music: Influences of proximal and distal cues and timbral characteristics. *Frontiers in Psychology, 4,* 789.

Tillmann, B., & Dowling, W. J. (2007). Memory decreases for prose, but not for poetry. *Memory and Cognition, 35,* 628–639. http://dx.doi.org/10.3758/BF03193301

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & D. W. Donaldson (Eds.), *Organization memory* (pp. 381–403). New York, NY: Academic Press.

Velardo, V., Vallati, M., & Jan, S. (2016). Symbolic melodic similarity: State of the art and future challenges. *Computer Music Journal, 40,* 70–83. http://dx.doi.org/10.1162/COMJ_a_00359

Ward, E. V., Berry, C. J., & Shanks, D. R. (2013a). Age effects on explicit and implicit memory. *Frontiers in Psychology*. Advance online publication. http://dx.doi.org/10.3389/fpsyg.2013.00639

Ward, E. V., Berry, C. J., & Shanks, D. R. (2013b). An effect of age on implicit memory that is not due to explicit contamination: Implications for single and multiple-systems theories. *Psychology and Aging, 28,* 429–442. http://dx.doi.org/10.1037/a0031888

White, B. W. (1960). Recognition of distorted melodies. *The American Journal of Psychology, 73,* 100–107. http://dx.doi.org/10.2307/1419120

Winold, A. (1975). Rhythm in twentieth-century music. In G. Wittlich (Ed.), *Aspects of twentieth-century music* (pp. 208–269). Englewood Cliffs, NJ: Prentice Hall.

Appendix

Stimuli

All stimuli as well as musical feature analysis of the melodies can be found in the online supplemental material S1 Stimuli.zip.

Experiments 1 and 2

The novel melodies used in Experiment 1 and 2 were composed in 12-TET. All melodies were 12 s in duration and unmistakably tonal. Half of the melodies were composed in major and the other half in minor. The meter was balanced across the melodies between 4/4 and 3/4. The tempi were pseudo-randomized between 80 and 165 beats per minute (bpm; $M = 120$ bpm). The rhythmic structure was kept simple with not more than two levels of metrical division (Winold, 1975). For more details about the stimuli, see Herff, Olsen, and Dean (2017), which is also the source of the following figure and the quote below.

Figure A1 shows representative examples of the melody corpus. An uninvolved expert listener with an extensive and sophisticated background in music (*Ollen Musical Sophistication Index* of 845, see Ollen, 2006, where on a scale of 0–1000, >500 is deemed to be musically sophisticated) described the melodies as follows:

"(. . .) I guessed they were theme tunes from TV programs, film music, or adverts. They sounded like the sort of melodies one would typically come across in everyday life."

Experiments 3 and 4

The melodies used in Experiment 3 and 4 originate from a large corpus of European folk songs. A thorough stimulus selection protocol was used to draw a representative sample of 98 folk melodies from the corpus. The stimulus selection procedure is detailed in Herff, Olsen, and Dean (2017), which is also the source of the figure below. Figure A2 shows representative examples of the melodies used in Experiments 3 and 4. The melodies had a mean duration of 10.84. All melodies were perceptually tested to be novel to a pilot sample of 12 Australian listeners.

Experiments 5 and 6

Experiment 5 and 6 used the same melodies as Experiments 3 and 4; however, the melodies were placed in a novel, unfamiliar tuning system. The system was designed to be an equally tempered tuning system that minimizes tonal affinity and similarity to 12-TET (Milne, 2013; Milne et al., 2011). The resulting system was 88.08 cents equal temperament. Using the tonal affinity model, this tuning system shows cosine similarity of .28058 to 12-TET (0 when the two systems are maximally dissimilar and 1 when they are identical, see Milne & Holland, 2016; Milne et al., 2015, 2016). Please refer to Herff, Olsen, Dean, et al. (2017) for further mathematical descriptions and perceptual tests of the tuning system. Besides the semi tone step-size (88.08 cents vs. 100 cents), the melodies in Experiment 5 and 6 were identical to those in Experiments 3 and 4. The Scale file of the 88.08-CET system can be found in the online supplemental material S1 Stimuli.zip.

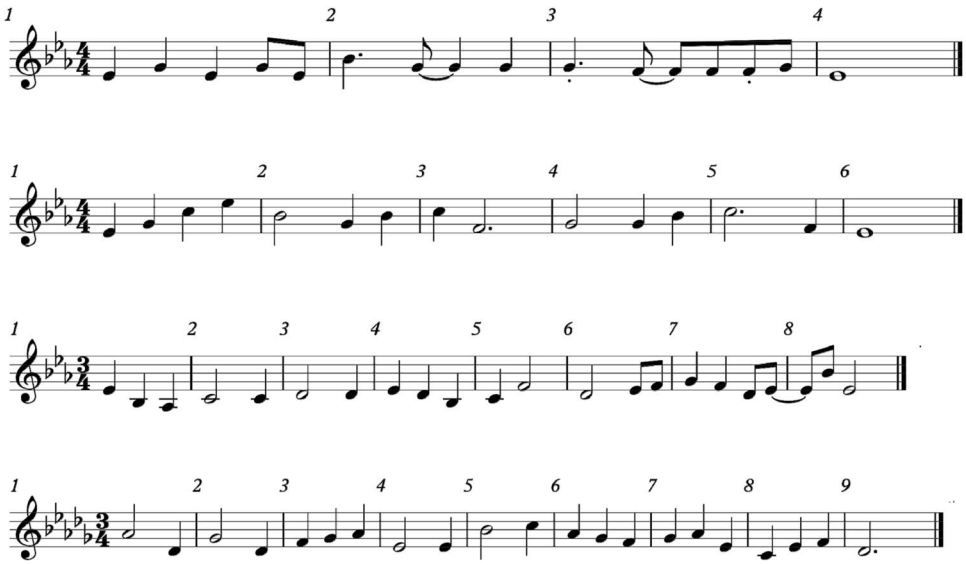

Figure A1. Examples of the melodies used in Experiments 1 and 2.

(*Appendix continues*)

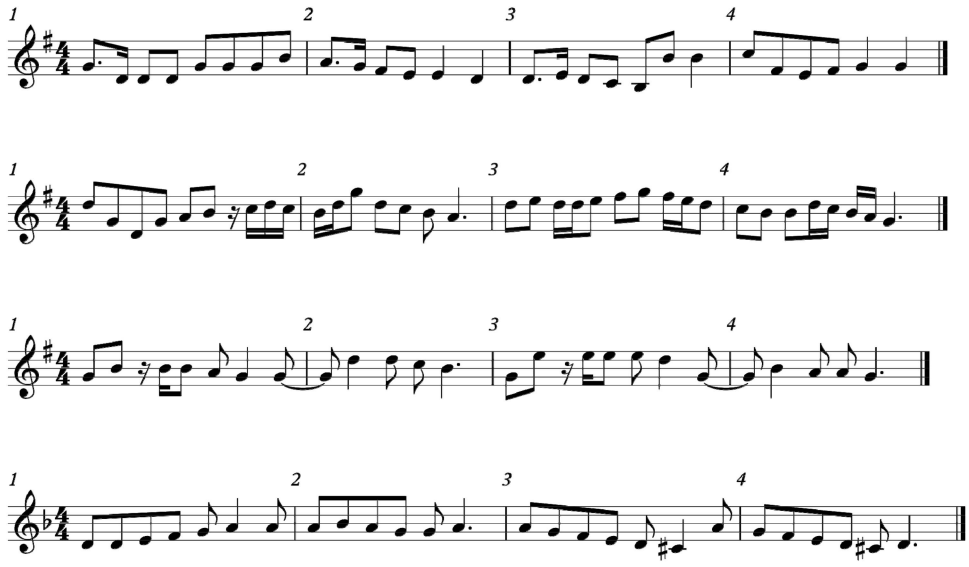

Figure A2. Representative examples of the melodies used in Experiments 3 and 4.

Experiment 7

The data from Experiment 7 are from a larger study conducted at Murdoch University, Australia, that investigated statistical learning of artificial pitch and rhythm grammars. The unfamiliar tuning system of Experiment 7 used five to six note melodies with the following pitch heights: 480, 520, 560, 605, and 665 Hz, as well as the following note durations: 60, 110, 550, and 920 ms, with a 100-ms silent gap between notes. All notes were synthesized pure tones with 10-ms linear onset and offset ramps. Durations and pitch discriminability were piloted to ensure that all pitch and duration differences were clearly discriminable ($N = 9$). The unequally tempered tuning system used in Experiment 7 is less similar to 12-TET than the tuning system used in Experiments 5 and 6. The tuning system used in Experiment 7 shows cosine similarity of .16186 to 12-TET and .22268 to the 88.08-CET system used in Experiments 5 and 6 (Milne, 2013; Milne et al., 2011). The above information and more information about the stimuli can be found in Herff, Olsen, Dean, et al. (2017) (Experiment 3).

Experiments 8 to 10

Experiment 8 uses the same stimuli as Experiment 7 described above. Experiment 9 used the same melodies as Experiment 8; however, all melodies are played with an isochronous

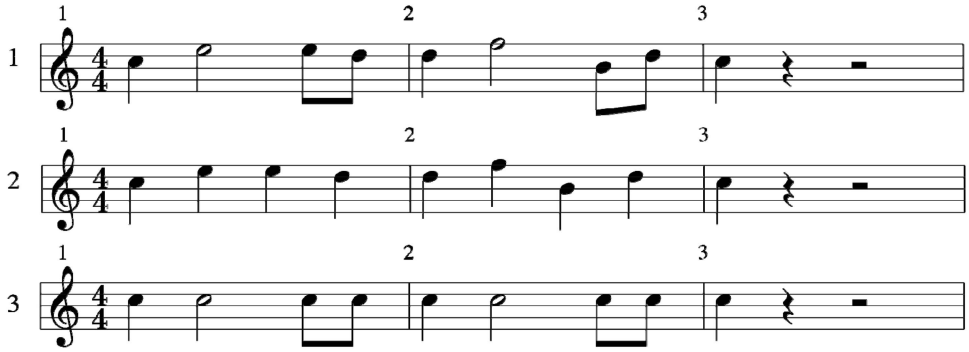

Figure A3. Example of the stimulus manipulations used in Experiments 8 to 10. Experiment 8 used melodies that consist of a combined melodic and rhythmic sequence (see 1). Experiment 9 used the pitch-only sequence of the original stimuli (see 2). Experiment 10 used the rhythm-only sequence of the original stimuli (see 3). Note that this figure is only an example of how the stimuli were manipulated. The actual stimuli presented in the study were melodies in an unfamiliar tuning system, and with more irregular note interonset intervals, as described in the Method section.

(*Appendix continues*)

rhythm, effectively removing rhythmic information between different melodies. This is illustrated in Figure A3, where Example 2 (similar to Experiment 9) is an isochronous (or pitch-only) version of Example 1 (similar to Experiment 8). Experiment 10 uses the same melodies as Experiment 8; however, all notes are played in the same pitch, effectively removing pitch information between different melodies. This is also illustrated in Figure A3, where Example 3 (similar to Experiment 10) is a same pitch (or rhythm-only) version of Example 1 (similar to Experiment 8). The figure originates from the work by Herff, Olsen, Prince, et al. (2017), where also further information about the stimuli can be found.

Received February 22, 2017
Revision received September 18, 2017
Accepted October 16, 2017 ■