Listeners perceive complex pitch-temporal structure in melodies

Jon B. Prince¹ · Shih En Jeanelle Tan¹ · Mark A. Schmuckler²

Published online: 9 December 2019
© The Psychonomic Society, Inc. 2019

Abstract

In typical Western music, important pitches occur disproportionately often on important beats, referred to as the tonal-metric hierarchy (Prince & Schmuckler, 2014, *Music Perception*, 31, 254–270). We tested whether listeners are sensitive to this alignment of pitch and temporal structure. In Experiment 1, the stimuli were 200 artificial melodies with random pitch contours; all melodies had both a regular beat and a pitch class distribution that favored one musical key, but had either high or low agreement with the tonal-metric hierarchy. Thirty-two listeners rated the goodness of each melody, and another 41 listeners rated the melodies' metric clarity (how clear the beat was). The tonal-metric hierarchy did not affect either rating type, likely because the melodies may have only weakly (at best) established a musical key. In Experiment 2, we shuffled the pitches in 60 composed melodies (scrambling pitch contour, but not rhythm) to generate versions with high and low agreement with the tonal-metric hierarchy. Both ratings of goodness ($N = 40$) and metric clarity ($N = 40$) revealed strong evidence of the tonal-metric hierarchy influencing ratings; there was no effect of musical training. In Experiment 3, we phase-shifted, rather than shuffled, the pitches from the composed melodies, thus preserving pitch contour. Both rating types (goodness $N = 43$, metric clarity $N = 32$) replicated the results of Experiment 2. These findings establish the psychological reality of the tonal-metric hierarchy.

Keywords Tonality · Meter · Tonal-metric hierarchy · Music cognition

Regardless of training, Western listeners immediately and automatically notice when they hear “wrong” notes or beats in a novel melody. This recognition occurs because musical key (tonality) and meter (beat structure) are fundamental components of music, evidenced by the rich literature on their psychological reality (for reviews, see Jones, 2016; Krumhansl & Cuddy, 2010). Tonality refers to the hierarchical arrangement of the 12 pitch classes (per octave) in the context of a musical key, and represents underlying psychological principles of cognitive reference points and sensitivity to statistical properties (Krumhansl & Cuddy, 2010). In the key of C major, the pitch classes C, E, and G are at the top level of the hierarchy, functioning as the most psychologically stable reference points in that key and heard most often. Regardless of training, Western listeners accept them as “good” notes and acceptable endpoints to melodies. In comparison, the pitch classes of D, F, A, and B form the middle level

of the hierarchy, and the notes C#, D#, F#, G#, and A# constitute the bottom level (rarely heard in a C major context). Meter refers to the regular alternation of strong and weak points in time (where the beat is), forming a hierarchical arrangement of points in time (London, 2012). The hierarchy consists of nested subdivisions of timespans (Lerdahl & Jackendoff, 1983), such that in the most common meter (4/4 time), the downbeat (start of the measure) is most stable, followed by the midpoint, then further bisections, and so on. When listeners tap their feet to a piece of music, they synchronize with the stable beats. Accordingly, the metric hierarchy is a central contributor to temporal organization in music, and forming expectations about event timing (Jones, 1976, 2016; Jones, Kidd, & Wetzell, 1981; Large & Palmer, 2002; Palmer & Krumhansl, 1990).

Although tonality and meter are complex hierarchies in the separate musical dimensions of pitch (Krumhansl, 1990) and time (Palmer & Krumhansl, 1990), they do not function in isolation—they are intertwined such that specific pitches occur at specific points in time. That is, typical Western music demonstrates a joint tonal-metric hierarchy (Prince & Schmuckler, 2014), such that pitches at the top of the tonal hierarchy occur disproportionately often on beats at the top of the metric hierarchy, and pitches with lower tonal stability are more likely to occur at points of lower metric stability. The evidence for the tonal-metric hierarchy comes from corpus studies that collapse

✉ Jon B. Prince
j.prince@murdoch.edu.au

¹ College of Science, Health, Engineering, and Education, Discipline of Psychology, Murdoch University, 90 South Street, Murdoch, WA 6150, Australia

² University of Toronto Scarborough, Toronto, Canada

across many pieces (Järvinen, 1995; Järvinen & Toivainen, 2000; Prince & Schmuckler, 2014); accordingly, its relevance at the level of individual pieces or melodies is unknown. Stated differently, is this joint hierarchy merely a theoretical curiosity, or are listeners actually sensitive to such co-occurrence in their processing of individual melodies?

The existing research on tonality and meter is mixed, with reports of these hierarchies functioning independently (Bigand, 1997; Palmer & Krumhansl, 1987a, 1987b; Prince, 2014a), interactively (Bigand, 1993; Boltz, 1989b, 1993, 1998; Schmuckler & Boltz, 1994), or asymmetrically—whereby tonality influences meter but not vice versa (Prince, Thompson, & Schmuckler, 2009) or the reverse (White, 2017)—depending on the specific musical context and stimuli, and experimental task. Parenthetically, by focusing here on tonality and meter, we leave aside the more general (and more complex) question of whether pitch and time are independent or interactive dimensions (for reviews, see Prince et al., 2009; Schellenberg, Stalinski, & Marks, 2014; Tillmann & Lebrun-Guillaud, 2006).

The research supporting independence of tonality and meter began with Palmer and Krumhansl (1987a, 1987b), who had participants rate the goodness or completion of different versions of a single musical sequence—either the pitch pattern (with isochronous timing), the rhythm pattern (with isotonic pitch), or the combined pitch and rhythmic patterns. These authors also used a phase-shift technique, whereby the original pitch and/or rhythmic pattern was shifted such that they began on a different element of the sequence (second, third, etc.). Linear combinations of the ratings from the pitch pattern condition and the rhythm pattern condition predicted the condition with both patterns, and interactive variables did not account for any variance. Using the same approach, Bigand (1997) reported similar results, predicting ratings with various pitch and temporal variables, including the tonal and metric hierarchies, but no interactive factors. Using a melodic similarity judgment while factorially manipulating the adherence to the pitch and temporal structure, Prince (2014a) found independent effects of tonality and meter, although introducing tempo changes altered the relative influence of pitch and temporal factors. Accordingly, there exists a considerable amount of work suggesting independence of tonal and metric processing of melodies.

However, there is also substantial work suggesting an interactive relationship between tonality and meter. Examining melodic closure ratings, Boltz (1989a, 1989b) found effects of both tonality and meter, along with an interaction—melodies that ended on time (consistent with the metric hierarchy) with the most stable pitch received disproportionately higher ratings than other combinations. Later work found that although detecting pitch changes was easier when it involved notes high in the tonal hierarchy, this benefit only occurred if such changes coincided with a temporal accent in a regular rhythm

(Boltz, 1993). The perception of sequence-final chords also showed interactive effects, whereby timing influenced belongingness judgments only when the chords belonged in the musical key of the sequence (Schmuckler & Boltz, 1994). This and other research have led to the conclusion that pitch and temporal factors are more likely to interact when they present a coherent joint structure, because they would reinforce one another, whereas incoherent melodies would foster more independent processing (Boltz, 1998). In contrast, goodness ratings of probe tones presented after a coherent melody (i.e., not evaluating the melody itself) can show asymmetric effects in which tonality interferes with metrical judgments but not vice versa (Prince et al., 2009). More recent research suggested that metric factors can influence the interpretation of the pitch structure, and not vice versa (White, 2017, but see Temperley, 2017).

Other investigations of tonal-metric relations have used artificial musical systems, which enable controlling carefully the numerous potentially confounding factors associated with naturalistic composed music. For instance, Rosenthal and Hannon (2016) exposed listeners to sequences with an unfamiliar distribution of pitch classes, and used metric structure to emphasize select pitches within the set. Both a matching task and a probe-tone rating task revealed strong influences of the metric stability on the apprehension of the novel pitch structure. Using a change-detection task, Acevedo, Temperley, and Pfordresher (2014) found that listeners were more likely to detect pitch changes when the metrical context coincided with the pitch pattern, supporting a metrical encoding hypothesis. Xin Wen and Krumhansl (2016) had listeners rate probe tones presented after each of the modes of the diatonic scale (i.e., using all possible starting points in the pitch interval pattern of 2 2 1 2 2 2 1 semitones). They also translated these patterns into durations and had listeners rate how well a dynamic (loudness) accent fit in all the different positions of the sequence. They correlated these ratings across dimension and found that the level of agreement across dimensions then predicted ratings in a subsequent experiment in which participants rated how well the pitch and durational pattern fit together, providing strong support for an interactive pitch–duration relationship.

Put together, there is a lack of clarity regarding how tonality and meter combine in music. Earlier research in this area used a small number (1–3) of unique melodies, and then created stimuli by shifting the pitch and temporal patterns out of phase (Bigand, 1997; Bigand & Pineau, 1996; Palmer & Krumhansl, 1987a, 1987b); it is possible that this repeated exposure to melodies influenced the findings. Other research used either judgments of probe tones following the melodies and/or sequences incompatible with (or highly atypical of) standard Western tonality (Rosenthal & Hannon, 2016; Xin Wen & Krumhansl, 2016). Most importantly, given that the existence of the tonal-metric hierarchy was only recently

established (Prince & Schmuckler, 2014), no research has investigated its role in melodic perception. We aimed to address this issue by testing whether listeners perceive melodies with aligned tonal and metric hierarchies (i.e., respecting the tonal-metric hierarchy) differently than melodies with misaligned hierarchies. Based on the research demonstrating untrained listeners' sensitivity to complex musical structure (for a review, see Bigand & Poulin-Charronnat, 2006), our basic prediction was that a lifetime of listening to Western music would provide sufficient passive exposure to internalize the joint tonal-metric hierarchy, such that listeners would be sensitive to violations of this structure.

It is also important to consider the potential role of surface-level information. Both tonality and meter are abstract structures—listeners must derive them from the patterns of pitch and timing that make up the musical surface, raising the possibility of complex interdependencies between surface and structure. In particular, pitch contour (the rising and falling across adjacent pitches that cumulatively shapes a melody) is a critical component of musical processing (Schmuckler, 2016) and may constrain the perception of tonality. For instance, Matsunaga and Abe (2005) found that changing the order of a pitch sequence altered which note listeners chose as the tonic (the central note of the musical key). Similarly, Anta (2015) found that identifying the tonic of a sequence was more difficult when the contour was distorted. Additionally, although Prince (2014a) observed largely independent effects of tonality, meter, contour, and rhythm, there were interactions between tonality and contour, such that changing the tonality of a melody only decreased its perceived similarity when the contour was kept the same. In the context of melodic memory, Dowling (1978) found that distinguishing two melodies with a similar contour was easier if they had a different tonal center. Therefore, in testing listeners' sensitivity to aligning the tonal and metric hierarchies, we incorporated manipulations of contour. Specifically, the stimuli of Experiments 1 and 2 were sequences with a random contour, whereas Experiment 3 preserved contour by phase-shifting a composed pitch pattern.

When not directing listeners' attention specifically to one dimension, pitch factors often dominate over temporal ones (Dawe, Platt, & Racine, 1993, 1994, 1995; Ellis & Jones, 2009; Hébert & Peretz, 1997; Prince & Loo, 2017; Prince et al., 2009). Accordingly, to ensure that our findings were not only reflecting pitch judgments, in both experiments we divided our participants into two groups; one group rated the sequences on overall melodic goodness, and the other group rated metric clarity (how clear the beat was). This manipulation provided convergent methods for assessing the role of the tonal-metric hierarchy. We predicted that listeners would give higher ratings of goodness and metric clarity to sequences with aligned tonal and metric hierarchies than those in which they were misaligned.

Experiment 1

Method

Participants The participants comprised a convenience sample of 73 undergraduate students at Murdoch University ($N = 28$) and the University of Toronto Scarborough ($N = 45$), who participated in exchange for course credit. Participants rated either goodness ($N = 32$) or metric clarity ($N = 41$) of the melodies. All participants reported normal hearing and having primarily listened to Western music throughout life. There were no prerequisites based on musical training, so the samples represented a normal cross-section of the undergraduate population. Unfortunately, the background questionnaires from the University of Toronto Scarborough students were lost, so the following demographic data are from the Murdoch students only. The average age was 26.6 years ($SD = 10.1$), the average years of formal musical training was 1.3 ($SD = 2.4$), and there were 19 female participants. The Murdoch University Human Research Ethics Committee approved the research (permit 2013/174, 2017/056), as did the University of Toronto Social Sciences, Humanities, and Education Research Ethics Board (Protocol 22701).

Materials Sequences were generated off-line, by scrambling the order of a preset distribution of pitch classes and metric positions, and then testing the alignment of the tonal and metric hierarchies. Every sequence had 33 notes distributed across four bars of the common 4/4 time signature (plus one final downbeat). The duration of each note was set to 125 ms, and each sequence lasted 8 s (tempo was 120 beats per minute). Table 1 shows the preset pitch class distribution (right column), which follows closely the Krumhansl and Kessler (1982) stability ratings for those pitches when presented in the context of a C-major musical key, $r(10) = .96$, $p < .001$. Table 2 shows the distribution of metric positions, which follows closely the Palmer and Krumhansl (1990) stability ratings for these metric positions in the context of the common 4/4 time signature, $r(14) = .90$, $p < .001$. In short, these sequences accorded strongly with the stability ratings of the major tonal hierarchy and the metric hierarchy of 4/4 time. Figure 1 shows the musical notation of two example sequences.

A custom script in MATLAB (Version 7.8; The MathWorks, Natick, MA) was used to create sequences by randomly scrambling the order of the pitches and assignment of metric positions to each bar. The final note was always the tonic (C), on the downbeat following the fourth bar. For each generated sequence, the script quantified the tonal stability and metric stability of each note, and then correlated these two vectors. We generated 100 “aligned” sequences with a minimum correlation of .33 ($M = .60$, $SD = .06$), and 100 “misaligned” sequences with a maximum average correlation

Table 1 Pitch class distribution of sequences in Experiment 1

Pitch class	Tonal stability	Occurrences
C	6.35	6
C#	2.23	1
D	3.48	3
D#	2.33	1
E	4.38	4
F	4.09	3
F#	2.52	1
G	5.19	6
G#	2.39	1
A	3.66	3
A#	2.29	1
B	2.88	3

Note. Tonal stability refers to the value from Krumhansl and Kessler (1982)

of .20 ($M = .05$, $SD = .12$). Misaligned sequences were easier to generate—when running the script without the restrictions specified above, a random sample of 500 sequences gave an average correlation of .05 ($SD = .17$). See Fig. 1 for an example of both types of sequence.

Using a piano timbre, sequences were converted to .wav files and presented to participants over Sennheiser HD280 Pro headphones. Custom scripts in MATLAB controlled the experimental interface, using the Psychophysics Toolbox (Brainard, 1997). All stimuli and raw data (for all

Table 2 Distribution of metric positions of sequences in Experiment 1

Metric position	Metric stability	Occurrences
0.00	6.10	5
0.25	4.00	1
0.50	4.20	2
0.75	3.33	1
1.00	5.95	4
1.25	3.00	1
1.50	4.35	2
1.75	3.45	1
2.00	5.75	4
2.25	2.80	1
2.50	3.05	2
2.75	3.15	1
3.00	5.45	4
3.25	3.30	1
3.50	4.70	2
3.75	4.30	1

Note. Metric position corresponds to when (in beats) the event occurs relative to the start of the bar; metric stability refers to the value from Palmer and Krumhansl (1990)

experiments) are available via the Open Science Framework (<https://osf.io/azxr2/>).

Procedure Participants provided informed consent and completed a background questionnaire (covering basic demographic information and musical experience). The experimenter explained the task to the participant, which for the goodness rating task was to “rate on a scale of 1 to 7 of how good, normal, or typical” the melody sounded. For participants completing the metric clarity task, instructions were instead to “rate on a scale of 1 to 7 how easy it is to find a regular beat in the melody.” For both tasks, participants were instructed not to rate their personal preference of the melody. The experimenter was present during four practice trials, to answer participant questions. Subsequently, each participant completed the full experiment (200 melodies) in a unique random order. The experimenter debriefed the participants afterward, and the entire process took about 1 hour.

Statistical analysis Ratings were averaged (within participant) separately for aligned and misaligned sequences. For both instructions in both experiments, a Bayesian ANOVA was used to test the difference between these values (the same pattern of results occurs when using a Frequentist ANOVA approach). Using G*Power (Version 3.1.9.2), we determined that our sample sizes were sufficient to detect an effect size as small as $\eta_p^2 = .019$ (goodness rating) and .015 (metric clarity rating) with alpha set to .05, and with power set to .95. A different analysis used correlation coefficients to explore whether the average interval size and number of reversals in each sequence predicted its rating (averaged across participant).

Results

The mean goodness ratings were 4.10 ($SD = .76$) and 4.12 ($SD = .77$) for the aligned and misaligned conditions, respectively. The Bayes factor (BF_{10}) was .277 (error 1.58%), indicating substantial support for the null model—that there was no difference in ratings between the two conditions. For the metric clarity task, the mean ratings were 4.52 ($SD = .85$) and 4.50 ($SD = .83$) for the aligned and misaligned conditions (respectively), giving a BF_{10} of .280 (error 2.59%), again giving substantial support for the null model.

Ratings were also averaged across participants to see how features of individual sequences influenced the data. In this item analysis, goodness ratings and metric clarity ratings of the 200 sequences correlated significantly with each other, $r(198) = .274$, $p < .001$. However, neither of the ratings correlated with the average pitch interval size (between adjacent notes), or the number of reversals in each melody (see Table 3); these two predictors (which previously have been shown to predict melodic perception; see Schmuckler, 1999,

Tonal-metric $r = .64$ (aligned)

Tonal-metric $r = .06$ (misaligned)

Fig. 1 Example sequences from Experiment 1, with an aligned (top) and misaligned (bottom) tonal-metric hierarchy. Both melodies have the same distribution of pitch classes and metric positions. All notes were 125-ms long, but for readability the notation filled silent gaps between notes

2009) were intercorrelated, $r(198) = .575$, $p < .001$. Because there was no relationship between the ratings and any of these predictors (including the tonal-metric alignment), there was no justification to further explore these factors using a regression analysis.

Discussion

The findings are clear—there was no effect of tonal-metric alignment on ratings of melodic goodness or metric clarity in Experiment 1. One explanation is that the randomly generated contours prevented the sequences from resembling typical music (i.e., violating norms of average pitch interval size and contour shape) and thus eliminated the more subtle effect of aligning the tonal and metric hierarchies. Normal melodies have small pitch intervals between adjacent notes; for instance, the average pitch interval of the 5,634 German folksongs of the Essen collection (Schaffrath, 1995) is 2.13 semitones ($SD = .49$). In contrast, the sequences of this experiment have an average pitch interval of 4.12 semitones ($SD = .39$). Similarly, there is a contour reversal every 4.17 notes ($SD = 2.87$) in the folksongs, compared with a reversal every 1.89 notes ($SD = .30$) in Experiment 1. It is possible that this characteristic interfered with establishing a tonal center, even though the sequences' pitch class distribution correlated with the tonal hierarchy of Krumhansl and Kessler (1982).

Another explanation of the null findings is that the sequences did not sufficiently resemble the pitch class distribution of normal melodies to establish a key center. Although the sequences adhered to the tonal hierarchy of Krumhansl and

Kessler (1982), this profile was derived from goodness-of-fit ratings of individual tones following a musical context, not a distribution of pitch classes in typical music. Each of the current sequences used all 12 pitch classes, but any given musical key includes only seven. Indeed, the German folksongs of the Essen collection use an average of 6.7 unique pitch classes ($SD = 1.0$), meaning that nondiatonic (out-of-key) pitches are exceedingly rare (see also Prince & Schmuckler, 2014). Therefore, using all five nondiatonic tones in each 33-note sequence (i.e., $M = 12$ pitch classes, $SD = 0$) may have prevented the activation of the tonal hierarchy in listeners. If so, then the metric alignment of the tones would be meaningless.

Although distributional statistics account for a remarkably large portion of the variation in key finding (Schmuckler & Tomovski, 2005), they are not the only factor (Temperley & Marvin, 2008). Indeed, recent research suggests that scattering pitches (as done here by randomizing pitch contour) degrades the tonal strength of melodies (Anta, 2015). If the sequences of the current experiment did not successfully establish a tonal center, the metric placement of pitch classes would be irrelevant (i.e., there can be no tonal-metric hierarchy without a tonal hierarchy). However, other research shows that participants perceive tonality in sequences with randomly generated contours, as long as the pitch class distribution aligns with the tonal hierarchy (Smith & Schmuckler, 2004). One relevant facet of the work by Smith and Schmuckler (2004) is that the duration or frequency of occurrence profiles of such randomly generated pitch sequences also needed to contain high differentiation between diatonic and nondiatonic tones (referred to by Smith & Schmuckler as the tonal magnitude of the sequence) to induce tonal percepts effectively. This idea supports the argument being developed here—that the sequences in this study might not have been sufficiently tonal for listeners.

It is also possible that the sequences did not evoke a strong metric framework, which would similarly invalidate the manipulation of aligning the tonal and metric hierarchies. However, this possibility seems particularly unlikely given that the distribution of note onsets corresponds quite closely

Table 3 Correlation of ratings (goodness, metric clarity) with interval size and reversals

	Metric clarity	Interval size	Reversals
Goodness	.27	−.09	.03
Metric clarity	—	−.02	.02
Interval size	−.02	—	.58

Note. Boldface values indicate significant correlations

with not only the metric hierarchy derived from goodness-of-fit ratings (Palmer & Krumhansl, 1990), but also with frequency of occurrence data. Specifically, the correlation coefficient between the distribution of note onsets in the Experiment 1 stimuli and the reported frequency of occurrence in a 4/4 time signature by Prince and Schmuckler (2014) is $r(14) = .98, p < .001$. Furthermore, listeners are prone to imposing a metric hierarchy on heard events—even isochronous sequences involuntarily invoke a metric hierarchy (Brochard, Abecasis, Potter, Ragot, & Drake, 2003; Potter, Fenwick, Abecasis, & Brochard, 2009). The goal of Experiment 2 was to determine which of the above explanations is most likely to account for the null findings of Experiment 1.

Experiment 2

Despite the existence of the tonal-metric hierarchy in typical Western music (Prince & Schmuckler, 2014), there was no evidence of listeners being sensitive to its presence in the artificial sequences of Experiment 1. This lack of an effect may have been due to the presence of nondiatonic tones disturbing the establishment of a key center, or the random nature of the pitch contour in the generated melodies. To determine which explanation is most likely, Experiment 2 used the same tasks, but generated stimuli by shuffling the pitch order of composed “source” melodies that rarely used nondiatonic tones. This shuffling operation meant that the stimuli still had random pitch contours, but left the rhythm intact. Furthermore, by using composed melodies, the stimuli gained a measure of ecological validity (admittedly reduced by scrambling the note order, and hence the pitch contour, but Experiment 3 addresses this issue). And finally, by using primarily diatonic tones these stimuli thus increased the difference in tone durations between diatonic and nondiatonic tones, consistent with Smith and Schmuckler’s (2004) notion of tonal magnitude.

We generated both an aligned and misaligned version of each source melody. We reasoned that if the use of nondiatonic tones was responsible for the null effect of Experiment 1, then an effect of tonal-metric alignment should emerge in Experiment 2. However, another null result would suggest either that the random pitch contour was to blame, or that listeners simply are not sensitive to the tonal-metric hierarchy.

Method

Participants Data were collected from University of Toronto Scarborough psychology undergraduate students in exchange for course credit. Forty participants (27 female) rated the goodness of the sequences, and another 40 (28 female) provided the metric clarity ratings. The average age was 19.1 years ($SD = 2.0$) and 18.7 ($SD = 0.9$), respectively; average

years of formal musical training was 2.1 ($SD = 3.3$) and 3.2 ($SD = 4.4$), respectively. Computer error resulted in the exclusion of the data from four additional participants.

Materials Sixty “source” melodies composed in the major mode and 4/4 meter were used to generate the 120 stimuli in this experiment. These source melodies had an average of 28.2 notes ($SD = 9.9$), and lasted on average 8.9 seconds ($SD = 1.2$), with a tempo of 120 beats per minute. Importantly, these melodies used an average of 6.85 unique pitch classes ($SD = 0.84$), making their pitch class distribution match that of typical Western music (in contrast to the sequences of Experiment 1 that each used all 12 pitch classes). Accordingly, the sequences more strongly evoked a tonal hierarchy. The average tonal-metric correlation of the source melodies was .22 ($SD = .23$). Although the source melodies were used for generating stimuli, they were not included in the experiment. Instead, a custom script in MATLAB generated two versions of each source melody—one with a high tonal-metric correlation ($M r = .65, SD = .05$) and one with a low tonal-metric correlation ($M r = -.63, SD = .03$). Figure 2 shows an example source melody and both the aligned and misaligned versions generated from it.

Because the source melodies were rhythmically diverse, randomizing the order of pitches introduced uncontrolled changes in the cumulative duration of each pitch class, which is the basis for quantifying the tonal strength of a melody (at least, via the distributional method; Krumhansl & Schmuckler, 1986; Takeuchi, 1994). Specifically, the average maximum key correlation (hereafter, key strength) for the aligned condition was 0.87 ($SD = 0.08$), compared with the misaligned condition mean of 0.63 ($SD = 0.19$); a Bayesian ANOVA gave extreme evidence of a difference between these two conditions, $BF_{10} = 8.2 \times 10^{+11}$ (error = 1.33%). This is because strong beats tend to feature longer notes, so putting tonally stable pitches on strong beats means that they will also be longer notes. We address this issue in the Results section. The melody conversion and the experimental interface were the same as in Experiment 1.

Procedure The procedure was the same as in Experiment 1, except that listeners rated 120 melodies instead of 200, and therefore the study took approximately 45 minutes. Additionally, because the random pitch shuffling meant that the pitch of the final note varied across trials (unlike Experiment 1), the instructions (both for goodness and metric clarity) were modified to explain that participants would hear a segment of a melody, which often would sound incomplete. All participants were instructed not to rate how well the melody ended, nor how much they liked it personally, but either how good, normal, or typical they thought the segment sounded (goodness rating) or how clear the pulse of the beat was (metric clarity rating).

(a) Source melody. Tonal-metric $r = .30$ (b) Aligned version. Tonal-metric $r = .61$ (c) Misaligned version. Tonal-metric $r = -.65$

Fig. 2 Three versions of an Experiment 2 melody. **a** The source melody, not used in the experiment. **b** A pitch-shuffled version with an aligned tonal and metric hierarchy. **c** A pitch-shuffled version with a misaligned tonal and metric hierarchy. Note that the rhythm is identical across versions

Statistical analysis The Bayesian statistical analysis approach was the same as Experiment 1. Had we used an ANOVA analysis, our sample sizes were sufficient to detect an effect size as small as $\eta_p^2 = .021$ in either instruction condition, at .95 power ($\alpha = .05$). We added regression analyses to assess the relative influence of key strength and the tonal-metric hierarchy in predicting ratings. In these analyses we also included a specification of the average tonal stability of the notes found in the melody, and a specification of the average metric strength (both measures are based on frequency of occurrence and therefore unaffected by note duration). In Experiment 1, all sequences were generated from the same pitch and metric distribution, which meant that every sequence had exactly the same average tonal stability and metric stability. In Experiment 2, however, the pitch and metric distribution could vary across source melody, and therefore might predict ratings. Nevertheless, because the tonal and metric stability measures were based on frequency of occurrence data, and both versions were generated from the same source melody, these measures were matched across aligned and misaligned versions of a given melody. That is, variations of average tonal stability and metric stability were completely independent from the manipulation of the tonal-metric hierarchy.

Results

Goodness ratings were 4.44 ($SD = .71$) and 4.15 ($SD = .76$) for the aligned and misaligned conditions, respectively, which provides extreme evidence for the tonal-metric hierarchy influencing ratings, $BF_{10} = 3367.61$ (error 1.43%). Metric clarity ratings were 4.61 ($SD = .84$) and 4.41 ($SD = .88$) for the aligned and misaligned conditions, which also provides

extreme evidence that the tonal-metric hierarchy influenced ratings, $BF_{10} = 130.07$ (error 1.30%). Ratings of goodness and metric clarity correlated at $r(118) = .75$, $p < .001$.

Because the pitch shuffling also changed the key strength of the sequences, in the item analysis we examined the relationship between ratings (averaged across participant), the tonal-metric correlation, the maximum key correlation, the average tonal stability (of all pitches in the sequence), and the average metric stability (of all metric positions in the sequence), in addition to the average pitch interval size and contour reversals. Again, the average tonal stability and average metric stability variables were identical for aligned and misaligned versions of a given melody, but varied across source melody.

Table 4 presents the regression equation for the goodness ratings, which showed significant contributions of the tonal-metric hierarchy, average tonal stability, average metric stability, and average size of adjacent pitch intervals. The key strength of the sequence did not contribute to the equation, nor did the number of reversals (standardized to melody length). That is, changes in key strength could not explain the positive effect on goodness ratings of aligning the tonal and metric hierarchies. The effects of tonal stability and metric stability occur because sequences with higher stability received higher ratings, which represents inherent differences between source melodies, independent from the effect of aligning the tonal and metric hierarchies. The effect of interval size shows that melodies with smaller intervals between adjacent notes received higher ratings, which makes sense considering that the average interval size was nearly twice that of typical melodies from the Essen collection (3.89 semitones vs. 2.13 semitones).

Table 4 Regression equation for goodness ratings of Experiment 2

	Beta (standardized)	<i>t</i> score	<i>p</i> value	<i>sr</i> ²	Tolerance
Tonal-metric correlation	.24	3.05	.003	.06	.99
Tonal stability	.25	2.84	.005	.05	.83
Metric stability	.20	2.27	.025	.03	.85
Interval size	-.28	-3.41	.001	.07	.97
<i>Key strength</i>	-.13	-0.80	.423	<.01	.33
<i>Reversal rate</i>	-.08	-0.79	.432	<.01	.67

Note. The bottom two rows (in italics) show the nonsignificant predictors as excluded variables (i.e., not included in the final equation) for reference. Squared semipartial correlations (*sr*²) indicate the variance uniquely attributable to that predictor. Total *r*² = .27

The regression equation for the metric clarity ratings is in Table 5, and is notably similar to the goodness ratings. Specifically, there were significant contributions of the tonal-metric hierarchy, average tonal stability, average metric stability, average pitch interval size, as well as no significant contribution from contour reversals. There was one difference from the goodness ratings, in that key strength was a significant predictor of metric clarity ratings. However, the relationship was counterintuitive—higher key strength was associated with *lower* ratings. This finding arguably disqualifies the contribution of key strength, especially given that the contribution of average tonal stability was both intuitive (greater stability led to higher ratings) and consistent (explained both goodness and metric clarity ratings). If nothing else, changes in key strength could not explain the observed effects of higher ratings for aligned melodies, because they would instead predict higher ratings for misaligned sequences.

To test the role of musical training in ratings, each participant's average rating for misaligned sequences was subtracted from their average rating for aligned sequences, and the resulting difference scores were correlated with participants' years of formal musical training. There were no significant relationships between training and the aligned/misaligned difference score, $r(38) = .267$, $p = .096$, and $r(38) = .077$, $p = .638$, for the goodness and metric clarity ratings, respectively. That is, formal training did not moderate the effect of aligning the tonal and metric hierarchies on ratings.

Discussion

By using composed melodies as the source material for generating stimuli in Experiment 2, we drastically reduced the number of nondiatonic pitches present in the sequences. Accordingly, we were able to uncover evidence of the tonal-metric hierarchy influencing ratings of both melodic goodness and metric clarity. Specifically, aligning the tonal and metric hierarchies led to higher goodness and metric clarity ratings. Although both the tonal-metric correlation and the distributional measure of key strength were higher for the aligned condition than the misaligned condition, only the tonal-metric correlation predicted the ratings. Tonal stability and metric stability also predicted the ratings, but as they were the same for both conditions, these factors cannot account for the main effect of aligning the hierarchies. The average size of adjacent pitch intervals also predicted both goodness and metric clarity ratings (smaller intervals led to higher ratings).

The predictions of the goodness ratings and metric clarity ratings showed high agreement overall, although there were some minor differences between rating tasks. In particular, average metric stability was a considerably stronger predictor of metric clarity than goodness, whereas interval size contributed more unique variance to goodness ratings than metric clarity. These findings align with the pattern that pitch factors dominate temporal factors in melodic judgments by default,

Table 5 Regression equation for metric clarity ratings of Experiment 2

	Beta (standardized)	<i>t</i> score	<i>p</i> value	<i>sr</i> ²	Tolerance
Tonal-metric correlation	.31	3.07	.003	.04	.43
Tonal stability	.33	3.71	<.001	.06	.59
Metric stability	.54	7.35	<.001	.24	.83
Interval size	-.18	-2.43	.017	.03	.85
Key strength	-.30	-2.52	.013	.03	.33
<i>Reversal rate</i>	-.07	-0.85	.398	<.01	.67

Note. The bottom row (in italics) shows the nonsignificant predictors as excluded variables (i.e., not included in the final equation) for reference. Total *r*² = .49

but that this pattern can reverse with appropriate participant instructions (Prince, 2011; Prince et al., 2009).

Comparing the current findings with the previous experiment reveals that the most likely explanation for the null result of Experiment 1 was that the sequences did not evoke a percept of a tonal hierarchy, despite following the pitch class distribution of the tonal hierarchy. Again though, as noted by Smith and Schmuckler (2004), simply aligning with a tonal pitch class distribution is not in and of itself sufficient to produce tonal percepts; instead, sufficient differentiation in durations between tones is necessary to induce such percepts. Accordingly, in Experiment 1 the alignment of the tonal and metric hierarchies had no effect on ratings because the prevalence of nondiatonic pitches prevented the activation of the tonal hierarchy. Comparing Experiments 1 and 2 also reveals that although scrambling pitch contour may interfere with establishing a tonal percept (Anta, 2015), it does not prevent it. Both experiments had a scrambled pitch contour, and yet one succeeded in demonstrating effects of the tonal-metric hierarchy whereas the other did not.

It is interesting that musical expertise did not correlate with sensitivity to aligning the tonal and metric hierarchies. Perhaps this finding should not be surprising, as long-term passive exposure is sufficient for listeners to develop sensitivity to sophisticated aspects of musical structure (Bigand & Poulin-Charronnat, 2006; Yates, Justus, Atalay, Mert, & Trehub, 2017). However, musicians tend to show stronger effects of the tonal hierarchy than untrained listeners, although its effects are clearly observable in both groups (Besson & Faïta, 1995; Halpern, Kwak, Bartlett, & Dowling, 1996; Krumhansl & Shepard, 1979). Similarly, for the metric hierarchy, the difference between musicians and untrained listeners is quantitative, not qualitative (Grahn & Rowe, 2009; Motz, Erickson, & Hetrick, 2013; Palmer & Krumhansl, 1990). As such, because the participants in the current studies were largely untrained, there was perhaps somewhat less sensitivity to the tonal hierarchy of the sequences (and more interference from the nondiatonic tones in Experiment 1). Put simply, when using musically untrained participants, activating the tonal hierarchy may require avoiding nondiatonic tones and employing pitch class distributions containing increased variance between the tone durations (i.e., a higher tonal magnitude).

Although these findings provide evidence for the psychological reality of the tonal-metric hierarchy, the sequences remain rather artificial by nature of their scrambled pitch contour. Furthermore, given the null effects of Experiment 1, it is worthwhile attempting to replicate the effects of tonal-metric alignment on listeners' percepts of musical sequences. Therefore, Experiment 3 used a convergent approach in which the pitch contour of each melody was phase-shifted to manipulate the tonal-metric alignment, as opposed to randomizing the pitch contour. This manipulation provides an opportunity

to test the effects of the tonal-metric hierarchy in the context of more naturalistic musical sequences (with an intact pitch contour), and replicate the effects observed in Experiment 2. It also enables a further test of the role of pitch contour—a stronger effect of the tonal-metric hierarchy in Experiment 3 (compared with Experiment 2) would provide evidence for a facilitatory role of intact pitch contour.

Experiment 3

In the final experiment, we created phase-shifted versions of the Experiment 2 source melodies (i.e., starting the melody on the n^{th} pitch instead of the first) to manipulate the tonal-metric hierarchy. This method preserves both the original pitch contour of the source melody and the original rhythm, such that only the alignment of the tonal and metric hierarchies differs. Although phase-shifting a melody is more ecologically valid than scrambling its contour, the process might introduce unanticipated decrements in its melodic goodness and metric clarity, quite apart from the tonal-metric manipulation. For instance, phase-shifting could disrupt the grouping and phrasing of the original melody. Therefore, we did not use any source melodies in their original (unshifted) form, because that would confound the presence of a phase shift with the tonal-metric manipulation. Instead, we applied two different phase shifts to each source melody, generating one version in which the tonal and metric hierarchies were aligned, and another version in which they were misaligned. Thus, any disruption of the grouping or phrasing of the original melody would affect both the aligned and misaligned versions, avoiding a confound with the tonal-metric manipulation. The details of the phase-shifting process are presented in the Method section. We predicted that the results would replicate those of Experiment 2, such that sequences with aligned tonal and metric hierarchies would receive higher goodness and metric clarity ratings than those with misaligned hierarchies.

Method

Participants There were 76 participants in Experiment 3, all from Murdoch University. For the participants rating goodness ($N = 44$, 35 female), the average age was 26.3 years ($SD = 9.8$), and the average years of formal musical training was 3.7 years ($SD = 4.9$). For the participants rating metric clarity ($N = 32$, 23 female), the average age was 28.7 years ($SD = 10.2$), and the average years of formal musical training was 1.7 ($SD = 2.4$). Recruitment and reimbursement procedures were the same as Experiment 1. Because of a computer error during data collection, the data of one participant (goodness rating) were excluded, so the final N for this group was 43.

Materials The 60 source melodies from Experiment 2 were again used to generate 120 (new) stimuli. Again, the source melodies were only used for generating stimuli and did not appear in the experiment; instead, a custom script in MATLAB was used to generate phase-shifted versions of these melodies. The relative pitch pattern was unaltered, but circularly shifted through every possible position except the original (i.e., starting the melody on the second, third, . . . , n^{th} note instead of the first). Meanwhile, the duration pattern remained in its original position. The versions (i.e., phase-shift lag values) that gave the highest and lowest tonal-metric correlation values were then used in the aligned and misaligned conditions, respectively, giving 120 stimuli in total. The average tonal-metric correlation for the aligned condition was .37 ($SD = .11$); for the misaligned condition it was $-.41$ ($SD = .13$). Figure 3 shows an example of a source melody as well as the aligned and misaligned versions generated from it.

As in Experiment 2, the rhythmic diversity of the melodies meant that the phase-shift manipulation introduced uncontrolled changes in the cumulative duration of each pitch class, affecting the measured tonal strength. The key strength for the aligned condition was .83 ($SD = .09$), compared with the misaligned condition mean of .75 ($SD = .12$); a Bayesian ANOVA gave extreme evidence of a difference between these two conditions, $BF_{10} = 847.60$ (error $<.00\%$). We return to this issue in the Results section. The melody conversion and the experimental interface were the same as in the previous experiments.

Procedure The procedure was the same as in Experiment 2.

Statistical analysis The Bayesian statistical analysis approach was the same as the previous experiments. Had we used an ANOVA analysis, our sample sizes were sufficient to detect an

effect size as small as $\eta_p^2 = .019$ (goodness rating) and .027 (metric clarity rating) at .95 power ($\alpha = .05$). The regression analyses were the same as in Experiment 2.

Results

Goodness ratings were 4.33 ($SD = .51$) and 4.20 ($SD = .49$) for the aligned and misaligned conditions, respectively, which provides substantial evidence for the tonal-metric hierarchy influencing ratings, $BF_{10} = 5.23$ (error 1.39%). Metric clarity ratings were 4.61 ($SD = .65$) and 4.45 ($SD = .58$) for the aligned and misaligned conditions, respectively, which provides very strong evidence that the tonal-metric hierarchy influenced ratings, $BF_{10} = 30.02$ (error 0.77%). Ratings of goodness and metric clarity correlated at $r(118) = .38$, $p < .001$.

For the regression analysis (as per Experiment 2), we excluded one sequence that had a rating (both for goodness and metric clarity) more than three standard deviations below the mean. Table 6 presents the regression equation for the goodness ratings, which showed significant contributions of the tonal-metric hierarchy, average tonal stability, and average metric stability (which had a counterintuitively negative relationship with ratings). The key strength of the sequence did not contribute to the equation, nor did either the number of reversals or the average size of adjacent pitch intervals. Thus, the positive effect of tonal-metric alignment on goodness ratings could not be explained by accompanying changes in key strength. As in Experiment 2, the effects of tonal stability and metric stability represent inherent differences between source melodies, independent from the effect of aligning the tonal and metric hierarchies.

The regression equation for the metric clarity ratings is in Table 7. There were significant contributions of the tonal-metric hierarchy, average tonal stability, average metric

(a) Source melody (phase shift lag = 0). Tonal-metric $r = .30$

(b) Aligned version (phase shift lag = -2). Tonal-metric $r = .51$

(c) Misaligned version (phase shift lag = -4). Tonal-metric $r = -.33$

Fig. 3 Three versions of an example melody. **a** The source melody, with no phase shift and not used in the experiment. **b** A phase-shifted version with an aligned tonal and metric hierarchy. **c** A phase-shifted version with

a misaligned tonal and metric hierarchy. The grey arrows in **b** and **c** indicate the original start point of the source melody

Table 6 Regression equation for goodness ratings of Experiment 3

	Beta (standardized)	<i>t</i> score	<i>p</i> value	<i>sr</i> ²	Tolerance
Tonal-metric correlation	.20	2.29	.024	.04	1.00
Tonal stability	.34	3.69	<.001	.10	.87
Metric stability	-.25	-2.77	.007	.06	.87
<i>Key strength</i>	.04	.37	.715	.00	.85
<i>Reversal rate</i>	.00	.00	.240	.00	1.00
<i>Interval size</i>	.02	.25	.803	.00	1.00

Note. The bottom three rows (in italics) show the nonsignificant predictors as excluded variables (i.e., not included in the final equation) for reference. Squared semipartial correlations (*sr*²) indicate the variance uniquely attributable to that predictor. Total *r*² = .16

stability, and average pitch interval size (smaller intervals were associated with higher metric clarity). As in the goodness ratings, neither the key strength nor the reversals predicted ratings. Thus, changes in key strength could not explain the observed effects of aligning the tonal and metric hierarchies on metric clarity ratings.

As in Experiment 2, there were no significant relationships between training and the aligned/misaligned difference score, $r(41) = .002$, $p = .989$, and $r(30) = .264$, $p = .145$, for the goodness and metric clarity ratings, respectively. That is, formal training again failed to moderate the effect of aligning the tonal and metric hierarchies on ratings.

Discussion

The sequences in Experiment 3 more closely resembled normal melodies than the previous two experiments (compare Fig. 3 to Figs. 1 and 2). Again, we discovered that adhering to the tonal-metric hierarchy increased ratings of both melodic goodness and metric clarity. We also replicated the finding that while the tonal-metric correlation predicted the ratings, key strength did not. Tonal stability and metric stability also predicted the ratings, but as they were the matched across versions, they could not account for the main effect of aligning the hierarchies. Interestingly, the size of the effect was larger in Experiment 2 than Experiment 3 (judging from both the mean rating differences and the Bayes factors). This change likely occurred because the range of the tonal-metric

manipulation was greater in Experiment 2 (aligned $r = .65$, misaligned $r = -.63$) than Experiment 3 (aligned $r = .37$, misaligned $r = -.41$). That is, larger differences between aligned and misaligned sequences (in terms of their adherence to the tonal-metric hierarchy) result in larger differences in ratings across these categories, in keeping with the regression results within each experiment.

There were again some differences between goodness and metric clarity ratings, in terms of which additional predictors contributed to the regression. Although the average tonal stability of the pitches in the melody (regardless of their metric placement) predicted both ratings, the average metric stability of the notes (regardless of pitch class) had an inconsistent relationship—specifically, a negative correlation with goodness and a positive correlation with metric clarity. Why? One possibility is that higher average metric stability is a proxy of syncopation—melodies with an abundance of temporal positions of high metric stability would not have many off-beat notes. Because moderate levels of syncopation contribute to groove and pleasure (Janata, Tomic, & Haberman, 2012; Madison, 2006; Witek, Clarke, Wallentin, Kringelbach, & Vuust, 2014), melodies with little to no syncopation would have high average metric stability, but also be less enjoyable than melodies with moderate syncopation, and therefore receive lower goodness ratings. However, when the task is to evaluate only how clear the beat is, low levels of syncopation would mean less potential confusion about beat placement, resulting in higher ratings of metric clarity. A related question

Table 7 Regression equation for metric clarity ratings of Experiment 3

	Beta (standardized)	<i>t</i> score	<i>p</i> value	<i>sr</i> ²	Tolerance
Tonal-metric correlation	.14	2.05	.043	.02	1.00
Tonal stability	.26	3.57	.001	.06	.87
Metric stability	.56	7.64	<.001	.26	.85
Interval size	-.18	-2.69	.008	.03	.96
<i>Key strength</i>	-.01	-1.16	.248	.00	.85
<i>Reversal rate</i>	-.12	-1.55	.125	.01	1.00

Note. The bottom two rows (in italics) show the nonsignificant predictors as excluded variables (i.e., not included in the final equation) for reference. Total *r*² = .49

is why this pattern occurred in Experiment 3, but not Experiment 2. The only difference between these experiments is that the stimuli in Experiment 3 kept the melodic contour intact, and perhaps this increased ecological validity affected how some of the musical characteristics influenced listeners' ratings. Of course, these explanations are speculative, and the safer conclusion is that average metric stability has an inconsistent relationship with ratings—or at least less consistent than average tonal stability and tonal-metric alignment.

Another difference between the goodness and metric clarity ratings was the role of interval size, which predicted only metric clarity ratings. Given that the interval patterns of the composed source melodies were unaltered (only phase-shifted) and therefore in the normal range ($M = 2.6$ semitones, $SD = .57$), there was no reason to think that average interval size would correlate with goodness. However, large pitch intervals function as accents, and as such contribute strongly to meter perception (Ellis & Jones, 2009; Hannon, Snyder, Eerola, & Krumhansl, 2004; Prince, 2014b). If the phase-shift moved pitch interval accents from their original position on strong beats and onto weak beats, metric clarity would decrease. Therefore, having fewer leaps overall (thus, a smaller average pitch interval) might improve metric clarity because it would prevent occasional weak beats from becoming accented by a large pitch interval and perceived as a strong beat.

As in Experiment 2, musical expertise was unrelated to sensitivity to tonal-metric alignment. Other research investigating the perception of both tonal and metric hierarchies (and using a phase-shift methodology) found that untrained listeners demonstrate sensitivity to these principles, although they also show stronger effects of more superficial information, such as pitch interval sizes (Bigand, 1993, 1997; Bigand & Pineau, 1996). However, in the present data, we observed no effect of musical training on the perception of an even more subtle statistical property than the tonal and metric hierarchies—namely, their alignment.

General discussion

In three experiments, we manipulated the alignment of the tonal and metric hierarchies in sequences that participants rated on goodness or metric clarity. There were no significant effects of this manipulation when the sequences included several nondiatonic tones (Experiment 1), but when composed melodies were used as source material (Experiments 2 and 3), both goodness and metric clarity ratings were higher for sequences with aligned tonal and metric hierarchies. These effects did not vary as a function of musical training, and could not be explained by concomitant variations in key strength. Accordingly, these data provide the first evidence of the psychological reality of the tonal-metric hierarchy.

Humans are adept at extracting and learning statistical regularities in the environment (Goldstone, 1998), including in a musical context (Loui, Wessel, & Hudson Kam, 2010; Prince, Stevens, Jones, & Tillmann, 2018; Rosenthal & Hannon, 2016; Selchenkova, Jones, & Tillmann, 2014; Tillmann, Stevens, & Keller, 2011). The tonal and metric hierarchies (separately) are good examples—in fact, they are characteristic features of Western music, and violating them leads to music that listeners generally reject as unpleasant (Ball, 2011; Prince, 2011). The tonal-metric hierarchy is another such regularity, so perhaps it is unsurprising that listeners were sensitive to it in the present experimental context. However, the tonal-metric hierarchy is different in that it emerges on the aggregate, often not describing individual musical segments. Nonetheless, the extent to which melodies conformed to this idealized joint distribution apparently influenced listeners' judgments of goodness and metric clarity.

Despite its psychological reality, the effects of the tonal-metric hierarchy rely on activating the perception of a musical key. That is, the tonal-metric hierarchy can only function when the sequences successfully establish a tonal center (i.e., Experiments 2 and 3). Interestingly, despite the central importance of melodic contour in perceptual organization in music (Bregman, 1990; Dowling, 1978; Schellenberg, 1996, 1997; Schmuckler, 2016), the tonal-metric hierarchy does not depend on having a typical melodic contour—Experiment 2 showed effects of the tonal-metric hierarchy despite having randomly shuffled pitch sequences. Furthermore, retaining an intact contour (Experiment 3) did not increase the effect size of the tonal-metric hierarchy relative to when the contour was random (Experiment 2).

These findings are an important extension to the literature. First and foremost, they establish the psychological reality of the tonal-metric hierarchy—it is more than a theoretical curiosity. Second, passive exposure to Western music is sufficient to demonstrate sensitivity to this statistical property, despite its subtle nature. Third, the tonal-metric hierarchy provides clear evidence for an interactive relation between the two hierarchies. This latter point provides an explanation for previous findings of independent contributions of tonality and meter. Although Prince (2014a) used normal melodies as source material, the factorial manipulations of respecting/violating surface and structural information in both pitch and time meant that the majority of the stimuli were not typical-sounding melodies (even more so than the melodies of the current Experiment 1). Other studies, instead of destroying the pitch and temporal patterns, used phase shifts similar to the current Experiment 3 (Bigand, 1997; Palmer & Krumhansl, 1987a, 1987b). A notable limitation of these previous studies is the familiarity effects from repeating the stimuli—Palmer and Krumhansl used a single melody, and Bigand used two. It is quite likely that over the course of an hour's exposure to the same underlying melody, participants were able to process the

stimuli in a more analytical manner, leading to more independent contributions of pitch and time (Jones & Boltz, 1989).

Each participant received one of two instructions: rate the overall goodness of the melody, or rate its metric clarity. The effects of tonal-metric alignment were consistent across both instructions in each experiment (both null in Experiment 1, both significant in Experiments 2 and 3). Furthermore, in Experiments 2 and 3, the average tonal stability of the notes in a sequence influenced the metric clarity rating. This finding adds to the evidence of tonality influencing the perception of metric information (Hannon et al., 2004; Prince, 2014b; Prince et al., 2009). However, it is also possible that participants rating metric clarity nonetheless incorporated pitch factors into their ratings (adventently or not). A more objective task, such as tapping along with the melody, could discriminate these possibilities. It also would provide an opportunity to replicate and extend the present findings regarding the role of the tonal-metric hierarchy.

Another issue to keep in mind is that the role of the tonal-metric hierarchy is almost certainly more nuanced than what can be observed with the current design. We used a rather blunt and categorical manipulation of the tonal-metric hierarchy in these experiments. Although aligning the tonal and metric hierarchies resulted in higher ratings, it is overly simplistic to think that good music would always follow the tonal-metric hierarchy. Indeed, the oscillation between tension and relaxation is a critical feature of good music (Krumhansl, 2015; Lerdahl & Jackendoff, 1983), and a continuous measure of the tonal-metric hierarchy (e.g., a moving-window average) may provide one such measure of this cycle. Thus, these data represent an important first step into testing the psychological reality of the tonal-metric hierarchy, but future research should focus on how more continuous variations in the alignment of the tonal and metric hierarchies might communicate musical emotion over the course of a piece of music. Additional areas of interest could include examining how the use of the tonal-metric hierarchy changes across musical styles and compositional periods. Despite the overall prevalence of the tonal-metric hierarchy in their corpus analysis, Prince and Schmuckler (2014) noted variations across composer. The perceptual consequences of such variation are unknown, although other work has shown that distributions of melodic intervals can be used to differentiate the major compositional styles in Western classical music (Zivic, Shifres, & Cecchi, 2013).

To summarize, we have provided the first evidence that the tonal-metric hierarchy described by Prince and Schmuckler (2014) has psychological relevance. Aligning the tonal and metric hierarchies resulted in higher ratings of both goodness and metric clarity. These findings advance our understanding of music cognition, pitch–time integration, and the role of perceptual organization in processing complex stimuli.

Open practices statement The stimuli and data for all experiments are available through the Open Science Framework (<https://osf.io/azxr2/>). None of the experiments were preregistered.

Author note Some of these data were included in an oral presentation at the International Conference on Music Perception and Cognition, July 2018. Partial support for this project was provided by an NSERC Discovery Grant provided to M. A. Schmuckler, and a Sir Walter Murdoch Distinguished Collaborator Award to J. B. Prince and M. A. Schmuckler. All stimuli and data are available through the Open Science Framework (<https://osf.io/azxr2/>)

References

- Acevedo, S., Temperley, D., & Pfordresher, P.Q. (2014). Effects of metrical encoding on melody recognition. *Music Perception*, 31, 372–386. doi:<https://doi.org/10.1525/mp.2014.31.4.372>
- Anta, J. F. (2015). Pitch: A key factor in tonality induction. *Music Perception*, 32, 413–433. doi:<https://doi.org/10.1525/mp.2015.32.4.413>
- Ball, P. (2011). Schoenberg, serialism and cognition: Whose fault if no one listens? *Interdisciplinary Science Reviews*, 36, 24–41. doi:<https://doi.org/10.1179/030801811x12941390545645>
- Besson, M., & Faïta, F. (1995). Event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 1278–1296. doi:<https://doi.org/10.1037/0096-1523.21.6.1278>
- Bigand, E. (1993). The influence of implicit harmony, rhythm and musical training on the abstraction of “tension-relaxation schemas” in tonal musical phrases. *Contemporary Music Review*, 9, 123–137. doi:<https://doi.org/10.1080/07494469300640391>
- Bigand, E. (1997). Perceiving musical stability: The effect of tonal structure, rhythm, and musical expertise. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 808–822. doi:<https://doi.org/10.1037/0096-1523.23.3.808>
- Bigand, E., & Pineau, M. (1996). Context effects on melody recognition: A dynamic interpretation. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 15, 121–134.
- 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, 100–130. doi:<https://doi.org/10.1016/j.cognition.2005.11.007>
- Boltz, M. G. (1989a). Perceiving the end: Effects of tonal relationships on melodic completion. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 749–761. doi:<https://doi.org/10.1037/0096-1523.15.4.749>
- Boltz, M.G. (1989b). Rhythm and good endings: Effects of temporal structure on tonality judgments. *Perception & Psychophysics*, 46, 9–17. doi:<https://doi.org/10.3758/BF03208069>
- Boltz, M. G. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception & Psychophysics*, 53, 585–600. doi:<https://doi.org/10.3758/BF03211736>
- Boltz, M. G. (1998). The processing of temporal and nontemporal information in the remembering of event durations and musical structure. *Journal of Experimental Psychology: Human Perception and*

- Performance*, 24, 1087–1104. doi:<https://doi.org/10.1037/0096-1523.24.4.1087>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436. doi:<https://doi.org/10.1163/156856897X00357>
- Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. Cambridge, MA: MIT Press.
- Brochard, R., Abecasis, D., Potter, D., Ragot, R., & Drake, C. (2003). The “ticktock” of our internal clock: Direct brain evidence of subjective accents in isochronous sequences. *Psychological Science*, 14, 362–366. doi:<https://doi.org/10.1111/1467-9280.24441>
- Dawe, L. A., Platt, J. R., & Racine, R. J. (1993). Harmonic accents in inference of metrical structure and perception of rhythm patterns. *Perception & Psychophysics*, 54, 794–807. doi:<https://doi.org/10.3758/BF03211804>
- Dawe, L. A., Platt, J. R., & Racine, R. J. (1994). Inference of metrical structure from perception of iterative pulses within time spans defined by chord changes. *Music Perception*, 12, 57–76. doi:<https://doi.org/10.2307/40285755>
- Dawe, L. A., Platt, J. R., & Racine, R. J. (1995). Rhythm perception and differences in accent weights for musicians and nonmusicians. *Perception & Psychophysics*, 57, 905–914. doi:<https://doi.org/10.3758/BF03206805>
- Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85, 341–354. doi:<https://doi.org/10.1037/0033-295X.85.4.341>
- Ellis, R. J., & Jones, M. R. (2009). The role of accent salience and joint accent structure in meter perception. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 264–280. doi:<https://doi.org/10.1037/a0013482>
- Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49, 585–612. doi:<https://doi.org/10.1146/annurev.psych.49.1.585>
- Grahn, J. A., & Rowe, J. B. (2009). Feeling the beat: Premotor and striatal interactions in musicians and nonmusicians during beat perception. *Journal of Neuroscience*, 29, 7540–7548. doi:<https://doi.org/10.1523/JNEUROSCI.2018-08.2009>
- Halpern, A. R., Kwak, S., Bartlett, J. C., & Dowling, W. J. (1996). Effects of aging and musical experience on the representation of tonal hierarchies. *Psychology and Aging*, 11, 235–246. doi:<https://doi.org/10.1037/0882-7974.11.2.235>
- Hannon, E. E., Snyder, J. S., Eerola, T., & Krumhansl, C. L. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, 30, 956–974. doi:<https://doi.org/10.1037/0096-1523.30.5.956>
- Hébert, S., & Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory & Cognition*, 25, 518–533. doi:<https://doi.org/10.3758/BF03201127>
- Janata, P., Tomic, S. T., & Haberman, J. M. (2012). Sensorimotor coupling in music and the psychology of the groove. *Journal of Experimental Psychology: General*, 141, 54–75. doi:<https://doi.org/10.1037/a0024208>
- Järvinen, T. (1995). Tonal hierarchies in jazz improvisation. *Music Perception*, 12, 415–437. doi:<https://doi.org/10.2307/40285675>
- Järvinen, T., & Toiviainen, P. (2000). The effect of metre on the use of tones in jazz improvisation. *Musicae Scientiae*, 4(1), 55–74. doi:<https://doi.org/10.1177/102986490000400103>
- Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, 83, 323–355. doi:<https://doi.org/10.1037/0033-295X.83.5.323>
- Jones, M. R. (2016). Musical time. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (2nd ed., pp. 125–141). New York, NY: Oxford University Press. doi:<https://doi.org/10.1093/oxfordhb/9780198722946.013.13>
- Jones, M. R., & Boltz, M. G. (1989). Dynamic attending and responses to time. *Psychological Review*, 96, 459–491. doi:<https://doi.org/10.1037/0033-295X.96.3.459>
- Jones, M. R., Kidd, G., & Wetzell, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1059–1073.
- Krumhansl, C. L. (1990). Cognitive foundations of musical pitch. New York, NY: Oxford University Press.
- Krumhansl, C. L. (2015). Statistics, structure, and style in music. *Music Perception*, 33, 20–31. doi:<https://doi.org/10.1525/MP.2015.33.1.20>
- Krumhansl, C. L., & Cuddy, L. L. (2010). A theory of tonal hierarchies in music. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Springer Handbook of Auditory Research: Vol. 36. Music perception* (pp. 51–87). New York, NY: Springer Science. doi:https://doi.org/10.1007/978-1-4419-6114-3_1
- Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89, 334–368. doi:<https://doi.org/10.1037/0033-295X.89.4.334>
- Krumhansl, C. L., & Schmuckler, M. A. (1986). *Key-finding in music: An algorithm based on pattern matching to tonal hierarchies*. Paper presented at the 19th Annual Meeting of the Society of Mathematical Psychology, Cambridge, MA.
- Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 579–594. doi:<https://doi.org/10.1037/0096-1523.5.4.579>
- Large, E. W., & Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science*, 26, 1–37. doi:[https://doi.org/10.1016/S0364-0213\(01\)00057-X](https://doi.org/10.1016/S0364-0213(01)00057-X)
- Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.
- London, J. (2012). *Hearing in time: Psychological aspects of musical meter*. Oxford, England: Oxford University Press.
- Loui, P., Wessel, D., & Hudson Kam, C. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music Perception*, 27, 377–388. doi:<https://doi.org/10.1525/mp.2010.27.5.377>
- Madison, G. (2006). Experiencing groove induced by music: Consistency and phenomenology. *Music Perception*, 24, 201–208. doi:<https://doi.org/10.1525/mp.2006.24.2.201>
- Matsunaga, R., & Abe, J. I. (2005). Cues for key perception of a melody: Pitch set alone? *Music Perception*, 23, 153–164. doi:<https://doi.org/10.1525/mp.2005.23.2.153>
- Motz, B. A., Erickson, M. A., & Hetrick, W. P. (2013). To the beat of your own drum: Cortical regularization of non-integer ratio rhythms toward metrical patterns. *Brain and Cognition*, 81, 329–336. doi:<https://doi.org/10.1016/j.bandc.2013.01.005>
- Palmer, C., & Krumhansl, C. L. (1987a). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 116–126. doi:<https://doi.org/10.1037/0096-1523.13.1.116>
- Palmer, C., & Krumhansl, C. L. (1987b). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, 41, 505–518. doi:<https://doi.org/10.3758/BF03210485>
- Palmer, C., & Krumhansl, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 728–741. doi:<https://doi.org/10.1037/0096-1523.16.4.728>
- Potter, D. D., Fenwick, M., Abecasis, D., & Brochard, R. (2009). Perceiving rhythm where none exists: Event-related potential (ERP) correlates of subjective accenting. *Cortex*, 45, 103–109. doi:<https://doi.org/10.1016/j.cortex.2008.01.004>

- Prince, J. B. (2011). The integration of stimulus dimensions in the perception of music. *Quarterly Journal of Experimental Psychology*, 64, 2125–2152. doi:<https://doi.org/10.1080/17470218.2011.573080>
- Prince, J. B. (2014a). Contributions of pitch contour, tonality, rhythm, and meter to melodic similarity. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 2319–2337. doi:<https://doi.org/10.1037/a0038010>
- Prince, J. B. (2014b). Pitch structure, but not selective attention, affects accent weightings in metrical grouping. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 2073–2090. doi:<https://doi.org/10.1037/a0037730>
- Prince, J. B., & Loo, L.-M. (2017). Surface and structural effects of pitch and time on global melodic expectancies. *Psychological Research/ Psychologische Forschung*, 81, 255–270. doi:<https://doi.org/10.1007/s00426-015-0737-y>
- Prince, J. B., & Schmuckler, M. A. (2014). The tonal-metric hierarchy: A corpus analysis. *Music Perception*, 31, 254–270. doi:<https://doi.org/10.1525/MP.2014.31.3.254>
- Prince, J. B., Stevens, C. J., Jones, M. R., & Tillmann, B. (2018). Learning of pitch and time structures in an artificial grammar setting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44, 1201–1214. doi:<https://doi.org/10.1037/xlm0000502>
- Prince, J. B., Thompson, W. F., & Schmuckler, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance* 35, 1598–1617. doi:<https://doi.org/10.1037/a0016456>
- Rosenthal, M. A., & Hannon, E. E. (2016). Cues to perceiving tonal stability in music: The role of temporal structure. *Music Perception*, 33, 601–612. doi:<https://doi.org/10.1525/mp.2016.33.5.601>
- Schaffrath, H. (1995). *The Essen folksong collection in kern format*. Retrieved from <http://essen.themefinder.org/>
- Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication realization model. *Cognition*, 58, 75–125. doi:[https://doi.org/10.1016/0010-0277\(95\)00665-6](https://doi.org/10.1016/0010-0277(95)00665-6)
- Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception*, 14, 295–318. doi:<https://doi.org/10.2307/40285723>
- Schellenberg, E. G., Stalinski, S. M., & Marks, B. M. (2014). Memory for surface features of unfamiliar melodies: Independent effects of changes in pitch and tempo. *Psychological Research*, 78, 84–95. doi:<https://doi.org/10.1007/s00426-013-0483-y>
- Schmuckler, M. A. (1999). Testing models of melodic contour similarity. *Music Perception*, 16, 295–326.
- Schmuckler, M. A. (2009). Components of melodic processing. In S. Hallam, I. Cross, & M. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 93–106). Oxford, England: Oxford University Press.
- Schmuckler, M. A. (2016). Tonality and contour in melodic processing. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (2nd ed., pp. 143–166). Oxford, England: Oxford University Press. doi:<https://doi.org/10.1093/oxfordhb/9780198722946.013.14>
- Schmuckler, M. A., & Boltz, M. G. (1994). Harmonic and rhythmic influences on musical expectancy. *Perception & Psychophysics*, 56, 313–325. doi:<https://doi.org/10.3758/BF03209765>
- Schmuckler, M. A., & Tomovski, R. (2005). Perceptual tests of an algorithm for musical key-finding. *Journal of Experimental Psychology: Human Perception and Performance*, 31, 1124–1149. doi:<https://doi.org/10.1037/0096-1523.31.5.1124>
- Selchenkova, T., Jones, M. R., & Tillmann, B. (2014). The influence of temporal regularities on the implicit learning of pitch structures. *Quarterly Journal of Experimental Psychology*, 67, 2360–2380. doi:<https://doi.org/10.1080/17470218.2014.929155>
- Smith, N. A., & Schmuckler, M. A. (2004). The perception of tonal structure through the differentiation and organization of pitches. *Journal of Experimental Psychology: Human Perception and Performance*, 30, 268–286. doi:<https://doi.org/10.1037/0096-1523.30.2.268>
- Takeuchi, A. H. (1994). Maximum key-profile correlation (MKC) as a measure of tonal structure in music. *Perception & Psychophysics*, 56(3), 335–346. doi:<https://doi.org/10.3758/bf03209767>
- Temperley, D. (2017). Commentary on Christopher White, “relationships between tonal stability and metrical accent in monophonic contexts”. *Empirical Musicology Review*, 12, 38–41.
- Temperley, D., & Marvin, E. W. (2008). Pitch-class distribution and the identification of key. *Music Perception*, 25, 193–212. doi:<https://doi.org/10.1525/mp.2008.25.3.193>
- Tillmann, B., & Lebrun-Guillaud, G. (2006). Influence of tonal and temporal expectations on chord processing and on completion judgments of chord sequences. *Psychological Research*, 70, 345–358. doi:<https://doi.org/10.1007/s00426-005-0222-0>
- Tillmann, B., Stevens, C., & Keller, P. E. (2011). Learning of timing patterns and the development of temporal expectations. *Psychological Research*, 75, 243–258. doi:<https://doi.org/10.1007/s00426-010-0302-7>
- White, C. (2017). Relationships between tonal stability and metrical accent in monophonic contexts. *Empirical Musicology Review*, 12, 19–37.
- Witek, M. A. G., Clarke, E. F., Wallentin, M., Kringelbach, M. L., & Vuust, P. (2014). Syncopation, body-movement and pleasure in groove music. *PLOS ONE*, 9. doi:<https://doi.org/10.1371/journal.pone.0094446>
- Xin Wen, O., & Krumhansl, C. L. (2016). *Isomorphism of pitch and time*. Paper presented at the the 9th International Conference of Students of Systematic Musicology (SysMus16), Jyväskylä, Finland.
- Yates, C. M., Justus, T., Atalay, N. B., Mert, N., & Trehub, S. E. (2017). Effects of musical training and culture on meter perception. *Psychology of Music*, 45, 231–245. doi:<https://doi.org/10.1177/0305735616657407>
- Zivic, P. H. R., Shifres, F., & Cecchi, G. A. (2013). Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 10034–10038. doi:<https://doi.org/10.1073/pnas.1222336110>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.