Farbood, M. M. & Mavromatis, P. (2018). The mutability of pitch memory in a tonal context. *Psychomusicology, 28*(1), 1-16.

The Mutability of Pitch Memory in a Tonal Context

Morwaread M. Farbood and Panayotis Mavromatis

New York University

Author Note

Morwaread M. Farbood, Department of Music and Performing Arts Professions, Steinhardt School, New York University. Panayotis Mavromatis, Department of Music and Performing Arts Professions, Steinhardt School, New York University.

Portions of this work were presented at the 12th International Conference on Music Perception and Cognition in Thessaloniki, Greece.

Correspondence should be addressed to Morwaread Farbood, Department of Music and Performing Arts Professions, 35 W. 4th St., Suite 1077, New York, NY 10012. E-mail: mfarbood@nyu.edu

Abstract

An experiment that investigates how a tonal context affects pitch recognition is presented. Melodic sequences that were composed to invoke varying degrees of tonality were rated by musicians ($N = 34$) for perceived strength of tonality. The sequences were then used in a pitch memory test based on a delayed-tone recognition paradigm. Listeners ($N = 48$) were asked to compare the first note of each melody (the standard) with a final, appended comparison tone that was either the same pitch or transposed by one semitone. The results showed that various factors including the presence of an interference tone one semitone away from the standard tone, the degree of tonality of the melodic sequence, and the tonal fitness of the standard and comparison tones predicted listener responses. In particular, the fitness of the comparison tone was a key factor in how listeners performed in the recognition task: comparison tones with higher fitness values increased performance when the comparison and standard were the same, but decreased performance when they were different. These results illustrate how tonality can both facilitate and interfere with pitch encoding and recognition, providing a detailed and definitive perspective on how pitch memory is influenced by tonal contexts.

Keywords: pitch, memory, tonality, music

The notion that tonality facilitates pitch memory in a musical context forms part of a larger narrative that has emerged from numerous studies exploring the influence of tonality on music processing. This broader perspective revolves around the concept of tonality as a hierarchical representation of music that provides an efficient encoding of pitch information. Studies examining the relationship between tonality and pitch recognition (Dewar, Cuddy, & Mewhort, 1977; Frankland & Cohen, 1996; Krumhansl, 1979; Long, 1977), melodic memory (Boltz, 1991; Cohen, Thorpe, & Trehub, 1987; Croonen, 1994; Cuddy, Cohen, & Mewhort, 1981; Cuddy, Cohen, & Miller, 1979; Cuddy & Lyons, 1981; Deutsch & Feroe, 1981; Dowling, 1978; Dowling, Kwak, & Andrews, 1995; Francès, 1988; Trainor & Trehub, 1992, 1993, 1994; Trehub, Cohen, Thorpe, & Morrongiello, 1986; Trehub, Schellenberg, & Kamenetsky, 1999; Watkins, 1985), and melodic similarity (Bigand, 1990; Dibben, 1994; Serafine, Glassman, & Overbeeke, 1989) all have all shown that a tonal context has a facilitating effect on pitch and melodic memory. However, in studies showing a general facilitating effect of tonality, negative effects have been occasionally noted, although the explanations for these effects have varied (Boltz, 1991; Dowling, 1991; Curtis & Bharucha, 2009; Frankland & Cohen, 1996; Krumhansl, 1979; Vuvan, Podolak, & Schmuckler, 2014). Some of these effects have been attributed to weak tonality or tonality "obscured by rhythmic accentuation" (Boltz, 1991). Others have been explained as errors due to differences in the stability or expectation of test tones in a strong tonal context. The current research offers a new perspective by directly investigating the negative effects of tonality on pitch memory while using more constrained stimuli than in prior work. Multiple factors both related and unrelated to tonality are also taken into account in the process of modeling the new data.

Methodological approaches used in experiments examining pitch memory in a musical context can be divided into two main categories: (1) a delayed-recognition task where individual pitches separated by interference tones are compared directly, and (2) single-pitch or whole-melody comparisons where target pitches are embedded within the sequences. In the first category, individual target pitches are specified and compared, while in the second category, the position of a target pitch within a sequence is unspecified—listeners are only aware that the target is one among a series of pitches in a given melody. In cases where comparisons involve multiple pitches, the comparison melody may be transposed as well.

This first methodological approach is exemplified in a series of studies by Wickelgren (1966, 1969) and Deutsch (1970, 1972a, 1972b, 1973a, 1973b, 1974, 1978). Wickelgren used the delayed recognition task in pitch memory experiments that featured sine-tone stimuli with an initial standard tone (a pitch that listeners were asked to remember) that was either fixed or variable in duration (2–8 s), followed by a single interference tone of variable length (0–180 s), then ending with a comparison tone that listeners were asked to judge as "same" or "different" from the standard tone. The results indicated that memory deterioration of the standard tone increased as the duration of the interference tone increased. Additionally, the decay rate of the memory trace appeared to be consistent regardless of the interference tone's intensity and frequency similarity (i.e., closeness) to the standard tone as well as the frequency difference between the standard and comparison tones.

While Wickelgren's studies featured only a single interference tone of variable length, Deutsch's experimental design used fixed durations for all tones while varying the number of interference tones, which ranged from four to eight (most commonly six). Deutsch also fixed the interonset intervals (IOIs) between the comparison and interference tones to 300 ms/200 beats per minute (BPM) and inserted a 2 s pause between the final interference tone and the comparison tone. Listeners were asked to judge whether the comparison tone was the same as the standard tone. In a series of experiments, Deutsch found that inserting various types of interference tones significantly affected performance. Adding an interference tone of the same pitch as the standard tone decreased error rate, while inserting an interference tone that was the same pitch as a nonmatching comparison tone had the opposite effect (Deutsch, 1972a). Deutsch (1972b) also manipulated the interference sequence so that the tone in the second serial position was a critical pitch—it was either the same as the standard tone or differed up to a whole tone in 1/6 tone increments. The results showed that error rates increased as the pitch distance of the critical tone increased, peaking at 2/3 of tone before falling rapidly as the distance increased to a whole tone. In a subsequent experiment, Deutsch (1973a) altered the second and third interference tones such that either one or both of them were a semitone lower or higher than the standard. Adding one such tone increased errors, and adding both pitches increased error even more.

In another set of experiments, Deutsch (1973b, 1974) examined whether inserting interference tones previously shown to increase error also affected performance when those tones were transposed by octaves. The results showed that when these critical pitches (e.g., one semitone from the standard or comparison) were transposed up or down an octave, they increased the error rate, and that this effect was stronger for the higher octave than the lower octave (Deutsch, 1973b). Transposition effects were further examined by comparing the effect of interference tones that were an octave higher or octave lower than the standard tone (Deutsch, 1974). All conditions caused interference, although there were more errors when all of the interference tones were in the higher octave as opposed to the lower octave. Moreover, when half of the tones were in each octave, the error rate was further increased. Deutsch (1978) also examined how contour affects pitch memory by manipulating six interference tones so that the intervals between the tones either increased or decreased monotonically or contained directional changes at least once every three intervals. Additionally, the tones were chosen either from one-octave or two-octave spans. The manipulation of these parameters resulted in conditions with varying average interval sizes. The results showed that errors increased as the average interval sizes increased. This effect of interval size as well as the earlier results showing the influence of repeated pitches and interference tones within a semitone of the standard were all considered in the data analysis that will be presented here.

Krumhansl (1979) used Deutsch's experimental paradigm to investigate the hypothesis that the psychological representation of a diatonic pitch in a tonal context is more stable than a nondiatonic pitch. Eight interference tones were used to either represent a tonal or atonal context. The tonal sequences consisted of all the notes of the C major scale and were designed to sound melodic; the atonal sequences were constructed by altering the lowest and highest pitches of the tonal sequences by a semitone. The standard tones were either diatonic or nondiatonic and designed to fall in the center of the sequence's pitch range. In the diatonic case, the tone was repeated in the interference

sequence. The IOIs between the interference tones were approximately 500 ms/120 BPM, and there was a 1.5 s pause before the final comparison tone. Listeners were asked to judge how similar the standard tone was to the comparison tone. The results indicated that diatonic standard tones were recognized better when the interference sequences were tonal, while nondiatonic tones were better remembered in atonal contexts. In a follow-up experiment, participants were asked to judge how "musical" the sequences sounded, resulting in ratings that matched the intended tonal/atonal categories.

Frankland and Cohen (1996) expanded on previous work by providing a more detailed account of how perceived tonality can affect pitch memory. Their hypothesis was that strength of key could predict performance on a pitch memory test. Their stimuli consisted of a standard tone followed by three interference tones and ending with a comparison tone, with 250 ms/240 BPM IOIs between all tones. The standard and comparison tones were always either C or C# and the interference tones consisted of all six sequential orders of the three tones of the pitches in C major, C minor, C# major, or C# minor triads. In a timed test, listeners were asked to indicate whether the test tones were the same or different. Key strength for each sequence was then used to calculate the stability of the standard tone ("Model 1") and the expectancy value of the comparison tone ("Model 2") in order to see if either value could predict performance. The results showed that both models predicted accuracy well, but only for musically trained listeners, and neither model was more effective than the other. Although not the direct focus of their study, the role of the comparison tone in Model 2 is the most relevant to the current work since it addresses the possibility that a tonal context, which relates directly to the expectancy of the comparison tone, is not beneficial for pitch recognition when pitches fail to align with that tonal context.

Frankland and Cohen (1996) and Krumhansl's (1979) experiments are the only studies that explore the effect of tonality using a delayed recognition task where *specified* target tones are compared directly. The second category of pitch memory studies employs a more indirect method: embedding the standard and comparison tones in melodic contexts. Dewar et al. (1977) explored the recognition of tones with and without melodic contexts. Listeners were presented with sequences of seven tones followed by a recognition test for one of the tones in the sequence. The serial position of the target tone varied and the IOIs between tones were approximately 670 ms/90 BPM. The probe tones consisted of either single tones or the full melodic context. Two sequence types, "musical" and "random," were tested; the former consisted of melodies taken from an ear-training book and the latter consisted of chromatic pitches randomly selected from an octave. The results indicated that pitch recognition was more accurate in the musical sequence condition, but better overall when embedded in a full melodic context (whether musical or random).

Long (1977) examined how various melodic features, including tonal strength, were correlated with pitch memory. The stimuli were taken from an earlier study by Taylor (1976), which investigated perceived tonality of melodies, since the tonality ratings of the sequences were already known. These 12 melodies, categorized as tonal or atonal, ranged from 7 to 15 pitches in length. The tones in the melodies formed palindromic pitch sequences (meaning they were identical played forwards or backwards) with isochronous IOIs of 500 ms/120 BPM. The standard tone was always the pitch adjacent to the center note of each sequence, and a 1 s comparison tone was appended to

each sequence following a 2 s pause. The comparison tones were either the same or one semitone lower/higher than the standard. The listeners' task was to determine whether the comparison tone occurred at some point in the melody or not. The results showed that pitches were recognized correctly more often when embedded in tonal rather than atonal sequences.

Vuvan et al. (2014) investigated how tonal expectancies influenced pitch recognition in a tonal context using a recognition memory task in which listeners heard a melody followed by a test tone and were asked to determine whether the tone matched any one of the previous pitches. They used major and minor folk song melodies as well as atonal melodies as stimuli. All melodies were played at a tempo of 120 BPM (500 ms IOI for quarter notes) and ranged from 7–8 s in total duration. They found that for tonal contexts, comparison tones with high tonal expectancies resulted in higher false recognition rates, while in atonal contexts, the lack of differentiated expectancies for tones resulted in no corresponding effect on pitch memory. In other words, a false-memory effect for pitch was demonstrated in a tonal context. Vuvan et al. framed these results in the context of the broader literature on false memory, schema theory, and the effects of availability and distinctiveness on memory. Although their goals were similar to those of the current study, the stimuli they used were far less constrained, featuring melodies with varied rhythmic patterns and repeated pitches.

The literature on melodic memory has been mentioned above in summary, but will not be detailed here since melodic recognition and pitch recognition are not equivalent tasks. However, one such study by Cuddy, Cohen, and Mewhort (1981) provides a useful template on how to design and analyze the relative tonality of melodic sequences. The intent of Cuddy et al. was to explore and identify the structural features that contributed to the perceived tonality of a melodic sequence. The goal was to develop a system for categorizing melodic sequences based on "formal rules of musical analysis and also the patterns of melodic contour." In a series of experiments, melodies designed to sound more and less tonal were constructed and evaluated. These evaluations included harmonic analyses from music theorists as well as ratings of "tonality or tone structure" from non-expert listeners. They proposed five clearly defined categories for harmonic structure ranging from the most tonal level, containing the notes of the major scale and an implied I-V-I progression, to the least tonal level, containing three non-diatonic tones. Sequences representing the five harmonic structure levels were then composed for a melodic memory test. Listeners were presented with a standard sequence and two transpositions, one of which was correctly transposed and the other of which had one note incorrectly transposed by one semitone. They were then required to choose which of the two transpositions was the correct one. Cuddy et al. found that their harmonic structure levels predicted accuracy and that complex contours combined with the least structure resulted in very poor performance.

In the present work, we attempt to replicate the results of earlier pitch memory experiments while adding constraints to the design of the stimuli. We then analyze this new data using generalized linear mixed-effects models (GLMMs) to predict listener judgments. We use Deutsch's experimental design as Krumhansl (1979) did; however, the sequences are designed in a musically rigorous manner similar to Cuddy et al. (1981). Beyond experimental design considerations, our analysis approach offers a more complete picture of how pitch memory is affected by tonal contexts.

Method

In order to design appropriate stimuli for a pitch memory experiment, melodic sequences of varying levels of tonality were first composed by the authors and then evaluated by musicians. Listeners were asked to rate the perceived tonality of the sequences, which were intended to range from atonal to very tonal. The goal was to then use these sequences—with a comparison tone appended—as stimuli for a pitch memory experiment modeled after Deutsch. The tonality ratings would provide a set of independent variables for the subsequent analysis of the data collected in the experiment. The protocols for both the ratings collection task and experiment were approved by the New York University institutional review board (University Committee on Activities Involving Human Subjects IRB# 10-0394).

Design and Evaluation of Melodic Sequences

Stimuli. There were a total of 60 melodic sequences composed, each consisting of seven pitches ranging from approximately one octave below to one octave above the standard tone. Only 11 sequences had ranges larger than an octave, and only three had ranges larger than a ninth. The sequences were constrained so that no pitch classes were repeated. Contour complexity of the sequences ranged from no changes in direction of intervals (i.e., a monotonically rising or falling contour) to a maximum of five directional changes. This resulted in an average number of direction changes across sequences of 2.63 (*SD* = 1.04) and an average interval size of 3.66 semitones (*SD* = 0.96). QuickTime MIDI grand piano timbre was used to render the sequences and the IOIs between note events were 600 ms (100 BPM). Each sequence was specifically designed to sound either tonal, weakly tonal, or atonal, with the assumption that the tonality ratings would span a fairly wide range. Tonal sequences clearly outlined chords arranged in an order that adhered to functional harmony. Weakly tonal sequences contained mostly diatonic pitch sets that did not outline harmonies. Atonal sequences contained pitch sets that were clearly nondiatonic. The purpose of composing the melodies within these categories was not to determine whether listeners agreed with those designations or not, but to ensure there would be a sufficiently wide range of perceived degrees of tonality represented in the stimuli for the pitch memory experiment.

Participants. Thirty-four musicians (26 males, 8 females) took part in the evaluation of the sequences with a mean age of 27.74 years (*SD* = 9.00). The majority were undergraduate and graduate students in a Psychology of Music course at New York University; three were music theory professors at various other institutions. Participants had an average of 12.53 years of formal training on a primary musical instrument (*SD* = 6.13) and an average overall musical training level (self-ranked) of 3.80 out of 5 (*SD* = 0.95). Two participants reported having absolute pitch.

Procedure. Participants navigated to a website where they were presented with a musical background survey followed by multiple-choice questions for each of the 60 melodic sequences, presented in random order. There was an embedded audio player that participants could click on to play the sequences. For the tonal-sounding melodies, each sequence was presented in a different transposition such that the key of a subsequent sequence was at least three steps away on the circle of fifths from the previous key. This

was done so that listeners would consider the tonal context of individual melodies separately. Participants were asked to provide three ratings: (1) how tonal each sequence sounded as a whole, (2) how tonal the first half of the sequence sounded, and (3) how tonal the second half sounded. They responded to each question by selecting one of five radio buttons, described as a continuum from "1 = Not tonal" to "5 = Clearly tonal." They were informed that they could play back the sequences repeatedly but were encouraged to listen carefully once and respond without over-thinking their answers. At the end of the study, they were asked whether they found themselves "hearing/labeling individual pitches as scale degrees" for sequences they felt were clearly tonal. The response choices ranged from "Not at all" (value = 1), through "Rarely", "Occasionally", and "Most of the time", to "Always" (value = 5). Lastly, there was a free response box where participants could enter additional comments.

Assessment of ratings. All of the sequences, in order of their overall tonal ratings, are shown Figure 1. The average whole-sequence tonality rating across all sequence means was 3.43 (SD = 1.10). The average first-half tonality rating was 3.52 (SD = 0.54), and the average second-half tonality rating was 3.52 (SD = 0.55). Although the first-half and second-half averages were practically identical, this was not reflected in the correlations between the mean ratings by sequence. There was a very strong correlation between the tonality ratings for the whole-sequence and the first half of the sequence with $r(58) = .85$, $p < .001$ and between the whole sequence and the second half with $r(58) = .80$, $p < .001$. However, the correlation between the first- and second-halves was weaker with $r(58) = .45$, $p < .001$.

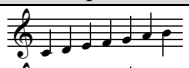
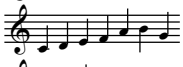
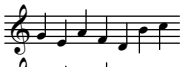
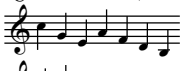
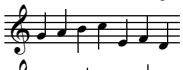
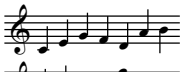
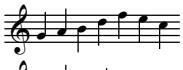
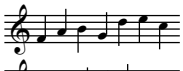
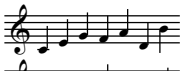
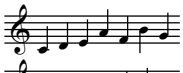
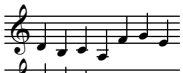
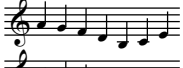
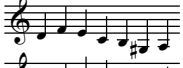
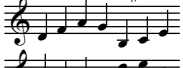
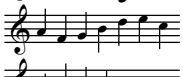
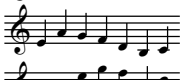
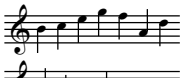
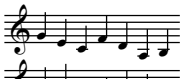
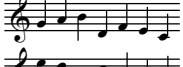
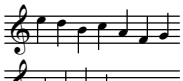
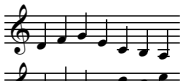
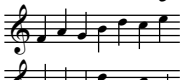
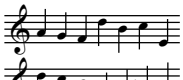
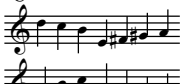
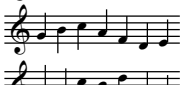
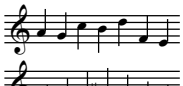
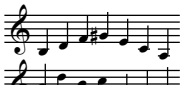
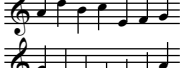
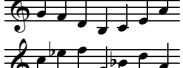
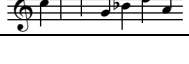
The average response to the question about identifying scale degrees was 3.53 (SD = 1.02). This indicated that most of the listeners were able to associate scale degrees with individual tones in a clearly tonal context. When looking at the participants' musical experience, scale degree identification ratings were more strongly correlated with overall musical background than music theory or instrumental background specifically. Table 1 shows the correlations between the various aspects of the participants' musical background (self-rated) and the scale-degree perception ratings. The correlation between music theory training and overall musical experience was considerably stronger than between instrumental training and overall experience; this probably reflects the nature of the participant pool, which included a sizable contingent of composers.

Table 1

Correlations between Musical Experience Ratings in Experiment 1.

	Scale Degree Identification	Theory Background	Instrumental Background
Theory Background	.48**		
Instrumental Background	.22	.28	
Overall Background	.59***	.77***	.25

Note. $df = 32$, *$p < .05$, **$p < .01$, ***$p < .001$.

Num	Notated Sequence	Ton All	Std All	Ton 1st	Ton 2nd	Cont Dir	Ave Int	Semi Pres	KK Stan	KKC Down	KKC Up
1		4.75	0.73	4.81	4.69	0	1.83	1	4.80	2.18	1.69
2		4.42	0.71	4.65	4.23	1	2.50	1	4.80	2.18	1.69
3		4.23	0.94	4.10	4.46	3	4.17	0	3.92	1.91	1.81
4		4.06	0.87	4.19	4.02	2	3.83	1	4.80	2.18	1.69
5		4.06	0.81	4.53	3.85	3	2.83	0	3.92	1.91	1.81
6		4.02	1.05	4.26	3.70	2	3.50	1	4.80	2.18	1.69
7		4.02	0.70	4.10	4.27	1	2.50	0	3.92	1.91	1.81
8		4.00	0.95	3.58	4.48	3	3.83	1	3.09	3.31	1.91
9		3.96	0.92	4.24	3.72	4	4.83	1	4.80	2.18	1.69
10		3.93	1.06	4.28	3.83	3	3.83	1	4.80	2.18	1.69
11		3.92	1.07	3.92	3.96	4	3.33	0	2.63	1.69	1.76
12		3.91	0.88	3.98	3.87	1	2.50	0	2.77	1.81	1.73
13		3.89	1.01	3.91	3.94	2	2.17	0	2.43	1.79	1.75
14		3.87	1.06	3.87	3.84	2	3.67	0	2.63	1.69	1.76
15		3.87	0.95	3.68	4.23	2	3.17	0	2.77	1.81	1.73
16		3.85	0.92	3.68	4.46	2	2.67	1	3.31	1.76	3.09
17		3.85	1.15	3.79	3.94	2	3.83	1	2.18	1.73	4.80
18		3.85	1.00	4.06	3.72	3	3.67	0	3.92	1.91	1.81
19		3.85	0.83	4.19	3.94	3	3.50	0	3.92	1.91	1.81
20		3.79	1.04	3.77	4.09	3	2.50	1	3.31	1.76	3.09
21		3.75	0.89	3.65	4.27	1	2.50	0	2.63	1.69	1.76
22		3.74	1.08	3.79	3.89	4	3.17	1	3.09	3.31	1.91
23		3.69	0.90	3.83	3.49	4	4.17	0	2.77	1.81	1.73
24		3.68	0.91	3.64	4.28	1	2.50	0	1.16	1.86	1.46
25		3.67	0.93	3.75	3.75	2	2.83	0	3.92	1.91	1.81
26		3.64	1.07	3.62	3.83	4	3.50	0	2.77	1.81	1.73
27		3.60	1.03	3.48	3.90	1	3.33	1	2.42	1.84	3.70
28		3.55	1.00	3.55	4.09	4	3.33	0	2.77	1.81	1.73
29		3.54	0.99	3.71	3.72	1	3.00	0	3.92	1.91	1.81
30		3.53	0.83	4.04	3.23	3	4.50	0	2.63	1.69	1.76

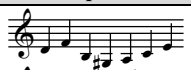
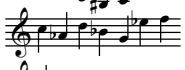
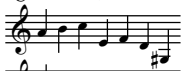
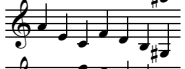
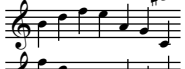
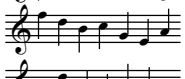
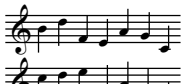
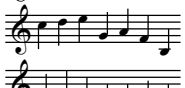
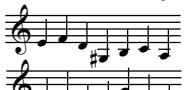
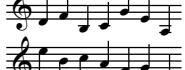
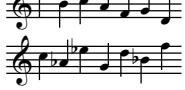
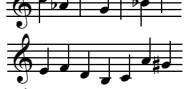
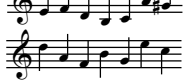
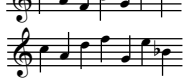
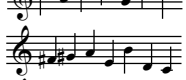
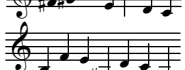
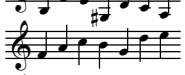
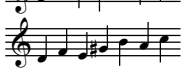
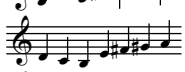
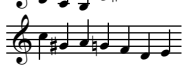
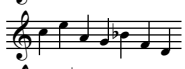
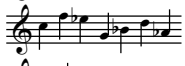
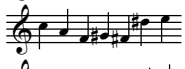
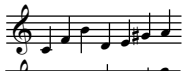
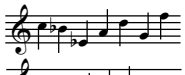
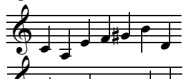
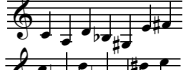
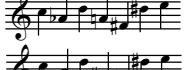
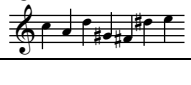
Num	Notated Sequence	Ton All	Std All	Ton 1st	Ton 2nd	Cont Dir	Ave Int	Semi Pres	KK Stan	KKC Down	KKC Up
31		3.52	0.99	3.31	3.94	2	3.33	0	2.43	1.79	1.75
32		3.52	1.07	3.69	3.63	3	4.50	0	3.31	1.76	3.09
33		3.48	0.98	3.89	3.11	3	3.50	1	4.35	2.18	1.84
34		3.47	1.23	3.51	3.45	2	3.83	1	4.35	2.18	1.84
35		3.46	0.92	3.44	3.65	1	3.83	1	2.18	1.73	4.80
36		3.45	1.08	3.70	3.64	3	3.33	1	3.09	3.31	1.91
37		3.41	1.07	3.39	3.78	3	4.50	1	2.18	1.73	4.80
38		3.40	1.11	4.10	2.96	3	4.17	1	4.80	2.18	1.69
39		3.38	1.16	3.10	3.46	3	2.83	1	3.26	1.75	2.73
40		3.36	1.09	3.53	3.37	3	4.50	0	2.63	1.69	1.76
41		3.35	0.91	3.71	3.19	4	3.33	1	3.31	1.76	3.09
42		3.32	1.02	3.49	3.57	5	6.17	0	4.80	2.18	1.69
43		3.30	1.06	3.57	3.09	3	3.00	1	3.26	1.75	2.73
44		3.29	0.85	3.21	4.02	4	5.33	0	2.63	1.69	1.76
45		3.15	1.05	3.56	2.65	4	6.00	0	3.92	1.91	1.81
46		3.11	0.91	3.89	2.77	3	4.33	0	1.76	1.13	1.18
47		3.10	1.21	2.85	3.42	3	4.33	1	2.42	1.84	3.70
48		3.06	1.07	3.00	3.52	2	3.50	1	3.09	3.31	1.91
49		3.06	1.05	3.00	3.26	4	2.67	0	2.43	1.79	1.75
50		3.04	1.13	3.38	3.29	1	2.17	0	1.16	1.86	1.46
51		3.00	0.83	2.83	3.57	3	2.33	0	4.38	1.99	1.54
52		2.98	0.91	3.31	3.40	3	4.00	0	3.92	1.91	1.81
53		2.93	0.98	3.34	2.87	3	4.67	0	4.80	2.18	1.69
54		2.89	1.11	2.96	3.32	3	3.67	0	2.27	1.48	1.09
55		2.85	1.03	2.43	3.60	2	4.50	1	3.70	2.42	1.79
56		2.75	1.04	2.85	2.67	3	6.17	0	2.63	1.69	1.76
57		2.67	1.06	3.19	2.34	2	4.33	1	3.70	2.42	1.79
58		2.60	0.83	3.27	2.31	3	4.00	0	2.02	1.32	0.98
59		2.53	1.04	2.36	3.35	3	4.67	0	1.91	1.25	0.92
60		2.50	0.88	2.69	3.08	3	4.33	0	1.91	1.25	0.92

Figure 1. (Shown on two prior pages) Sequences (without comparison tones appended) ordered by whole-sequence tonality ratings. Ton All: mean overall tonality rating; Std All: standard deviation of the overall tonality rating; Ton 1st: mean first-half tonality rating; Ton 2nd: mean second-half tonality rating; Cont Dir: number of contour direction changes; Ave Int: mean interval size in semitones; Semi Pres: a pitch one semitone away from the standard tone (or octave equivalent) is present in interference tones; KK Stan: fitness value of the standard, based on Krumhansl-Kessler key-profile value (also identical to the fitness of the comparison tone in the "same" condition); KKC Down: fitness value of the comparison tone one semitone lower than the standard; KKC Up: fitness value of the comparison tone one semitone higher than the standard.

The general comments provided by the listeners did not converge on any particular issues. One participant commented that it was difficult to perceive any of the sequences as atonal, while two others felt that many of the sequences sounded atonal, at least initially. Another participant commented that number of changes in contour were more important than a sense of scale degree. There were two comments on the difficulty of rating the two halves separately due to the lack of clarity on how to determine the midpoint boundary. One listener remarked that at times the two halves sounded independently tonal, but when combined, the sequence as a whole sounded atonal due to an apparent key change in the middle. Despite those comments, it appeared that participants did not have too much trouble with the task, particularly when judging the sequences as a whole.

Pitch Memory Experiment

Stimuli. The stimuli for the pitch memory experiment consisted of the 60 melodic sequences with a comparison tone added at the end of each sequence (note that term "sequence" as used here and in subsequent discussions refers only to the original seven-note melodies and does not include the comparison tone). QuickTime MIDI grand piano timbre was used to render the stimuli, which were then converted to 16-bit, 16 kHz mono WAV files for playback. The IOIs between all of the tones in each sequence were 600 ms (100 BPM), with a 1200 ms silence inserted between the last note of the sequence and the comparison tone. The first pitch of the sequence was designated the standard tone. The comparison tone was the same pitch as the standard tone, one semitone higher, or one semitone lower. Although restricting the comparison tone to this narrow pitch range resulted in unbalanced comparison tone types (i.e., different numbers of matching/nonmatching and within-key/out-of-key comparison tones), this constraint was deemed necessary in order to avoid having pitch distance be a potentially confounding factor.

Participants. Forty-eight participants took part in the pitch memory experiment, mean age 23.65 years (SD = 6.47), 22 female, 26 male. There was no overlap with the pool of listeners who provided the tonality ratings for the sequences. Both musicians and nonmusicians were recruited, although all but one participant had at least some minimal musical training. The mean number of years of formal training on a primary instrument was 9.88 years (SD = 5.46) and the mean overall musical training level (self-ranked) was 3.43 out of 5 (SD = 1.07). Five participants reported having absolute pitch.

Procedure. Participants were asked to fill out a musical background survey before being seated in front of a computer, where they listened to each stimulus over Sennheiser HD650 headphones in a sound isolated (hemi-anechoic) chamber. Each

listener was presented with the 180 stimuli (60 sequences × 3 possible comparison tones) *twice*, with the exception of the first two participants, who heard each stimulus three times. The number of repetitions was reduced because it quickly became apparent the experiment was too long. The stimuli were presented in pseudorandom order such that no trial was preceded by another trial that featured a stimulus with the same comparison tone type or the same sequence. Furthermore, the stimuli were randomly transposed to a new key for each trial, ranging from 11 semitones lower to 11 semitones higher than the original key (with lower/higher transpositions alternating every trial). The experiment was presented on a MATLAB graphical interface that used Psychtoolbox3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) for audio playback. Participants pressed a "Play sound" button to hear each stimulus once, after which the button was grayed out to prevent further playback. They were asked to determine whether the first pitch of each stimulus was the same or different from the final pitch they heard. Listeners responded by selecting one of two radio buttons labeled "Same" or "Different." The experiment took approximately one hour to complete.

Results

Overall performance was first examined by looking at mean percent-correct values and d' values. Table 2 shows the mean performance for each sequence broken down by comparison-tone type. From a general perspective, listener performance was worse on trials where the comparison tone was different from the standard tone, suggesting a "same" response bias. Mean response (percent correct) was 70.97% for trials with different comparison tones and 86.67% for trials with same comparison tones. Across participants, the average percent-correct score was 76.21% (SD = 15.30, min = 44.17, max = 97.50) and the mean d' value was 1.91 (SD = 1.03, min = -0.03, max = 3.80). In cases where the hit rate and false-alarm rate were either 0 or 1 when calculating d' values, the "loglinear" approach (Hautus, 1995) was used to avoid extreme values: that is, 0.5 was added to the number of hits and false alarms, and 1 was added to the number of signal and noise trials. The five listeners who reported having absolute pitch (AP) performed better than the non-AP participants, mean 85.33% versus 75.14% correct responses, mean d' of 2.51 versus 1.84. Since AP participants' performance was far from perfect (and did not include the highest score of 97.50%), their data were retained in the subsequent analyses.

The correlation between the mean whole-sequence subjective tonality ratings and mean percent-correct values by sequence was high, $r(58) = .68$, $p < .001$. This indicated a strong positive relationship between the perceived tonality of the sequences and listeners' performance on the pitch memory task, reflecting the conclusions reached in earlier research. However, as further analysis reveals, these results do not represent the full effects of tonality in the context of the pitch memory task.

Evaluating Tonal Context

The first step in the main analysis required determining the key of each sequence to calculate other metrics. These metrics were then used as independent variables in a generalized linear mixed-effects model. Similar to Frankland and Cohen's (1996)

Table 2

Independent Variables and Mean Performance for Each Melodic Sequence.

Seq Num	Ton All	Ton 1st	Ton 2nd	Cont Dir	Ave Int	Semi Pres	KK Stan	KKC Down	KKC Up	Same Resp (%)	Down Resp (%)	Up Resp (%)
1	4.75	4.81	4.69	0	1.83	1	4.80	2.18	1.69	84.7	86.7	88.8
2	4.42	4.65	4.23	1	2.50	1	4.80	2.18	1.69	84.8	79.8	85.9
3	4.23	4.10	4.46	3	4.17	0	3.92	1.91	1.81	93.9	79.6	61.2
4	4.06	4.19	4.02	2	3.83	1	4.80	2.18	1.69	79.6	74.5	86.7
5	4.06	4.53	3.85	3	2.83	0	3.92	1.91	1.81	90.9	82.8	72.7
6	4.02	4.26	3.70	2	3.50	1	4.80	2.18	1.69	91.8	84.7	81.6
7	4.02	4.10	4.27	1	2.50	0	3.92	1.91	1.81	92.9	85.7	82.7
8	4.00	3.58	4.48	3	3.83	1	3.09	3.31	1.91	83.7	77.6	80.6
9	3.96	4.24	3.72	4	4.83	1	4.80	2.18	1.69	91.8	87.8	85.7
10	3.93	4.28	3.83	3	3.83	1	4.80	2.18	1.69	83.7	84.7	80.6
11	3.92	3.92	3.96	4	3.33	0	2.63	1.69	1.76	89.8	66.3	85.7
12	3.91	3.98	3.87	1	2.50	0	2.77	1.81	1.73	91.9	81.8	83.8
13	3.89	3.91	3.94	2	2.17	0	2.43	1.79	1.75	93.9	61.2	69.4
14	3.87	3.87	3.84	2	3.67	0	2.63	1.69	1.76	87.8	78.6	77.6
15	3.87	3.68	4.23	2	3.17	0	2.77	1.81	1.73	82.7	78.6	76.5
16	3.85	3.68	4.46	2	2.67	1	3.31	1.76	3.09	92.9	75.5	46.9
17	3.85	3.79	3.94	2	3.83	1	2.18	1.73	4.80	75.5	76.5	32.7
18	3.85	4.06	3.72	3	3.67	0	3.92	1.91	1.81	90.8	72.4	86.7
19	3.85	4.19	3.94	3	3.50	0	3.92	1.91	1.81	90.8	71.4	76.5
20	3.79	3.77	4.09	3	2.50	1	3.31	1.76	3.09	87.8	76.5	77.6
21	3.75	3.65	4.27	1	2.50	0	2.63	1.69	1.76	89.8	76.5	73.5
22	3.74	3.79	3.89	4	3.17	1	3.09	3.31	1.91	70.4	76.5	73.5
23	3.69	3.83	3.49	4	4.17	0	2.77	1.81	1.73	90.8	79.6	71.4
24	3.68	3.64	4.28	1	2.50	0	1.16	1.86	1.46	87.8	76.5	82.7
25	3.67	3.75	3.75	2	2.83	0	3.92	1.91	1.81	90.8	74.5	76.5
26	3.64	3.62	3.83	4	3.50	0	2.77	1.81	1.73	88.8	70.4	66.3
27	3.60	3.48	3.90	1	3.33	1	2.42	1.84	3.70	82.8	68.7	64.6
28	3.55	3.55	4.09	4	3.33	0	2.77	1.81	1.73	86.7	79.6	69.4
29	3.54	3.71	3.72	1	3.00	0	3.92	1.91	1.81	93.9	76.5	86.7
30	3.53	4.04	3.23	3	4.50	0	2.63	1.69	1.76	94.9	67.3	74.5
31	3.52	3.31	3.94	2	3.33	0	2.43	1.79	1.75	87.8	79.6	83.7
32	3.52	3.69	3.63	3	4.50	0	3.31	1.76	3.09	86.7	80.6	68.4
33	3.48	3.89	3.11	3	3.50	1	4.35	2.18	1.84	93.9	63.3	70.4
34	3.47	3.51	3.45	2	3.83	1	4.35	2.18	1.84	87.8	76.5	82.7
35	3.46	3.44	3.65	1	3.83	1	2.18	1.73	4.80	86.7	56.1	55.1
36	3.45	3.70	3.64	3	3.33	1	3.09	3.31	1.91	79.6	54.1	81.6
37	3.41	3.39	3.78	3	4.50	1	2.18	1.73	4.80	78.6	60.2	59.2
38	3.40	4.10	2.96	3	4.17	1	4.80	2.18	1.69	87.8	60.2	63.3
39	3.38	3.10	3.46	3	2.83	1	3.26	1.75	2.73	93.9	69.4	72.4
40	3.36	3.53	3.37	3	4.50	0	2.63	1.69	1.76	83.7	70.4	77.6
41	3.35	3.71	3.19	4	3.33	1	3.31	1.76	3.09	90.8	82.7	73.5
42	3.32	3.49	3.57	5	6.17	0	4.80	2.18	1.69	88.8	87.8	68.4
43	3.30	3.57	3.09	3	3.00	1	3.26	1.75	2.73	92.9	61.2	57.1
44	3.29	3.21	4.02	4	5.33	0	2.63	1.69	1.76	89.8	70.4	76.5
45	3.15	3.56	2.65	4	6.00	0	3.92	1.91	1.81	82.7	69.4	61.2
46	3.11	3.89	2.77	3	4.33	0	1.76	1.13	1.18	88.9	56.6	59.6
47	3.10	2.85	3.42	3	4.33	1	2.42	1.84	3.70	84.7	71.4	50.0
48	3.06	3.00	3.52	2	3.50	1	3.09	3.31	1.91	90.8	69.4	74.5
49	3.06	3.00	3.26	4	2.67	0	2.43	1.79	1.75	80.6	84.7	68.4
50	3.04	3.38	3.29	1	2.17	0	1.16	1.86	1.46	82.7	65.3	67.3
51	3.00	2.83	3.57	3	2.33	0	4.38	1.99	1.54	88.9	70.7	87.9
52	2.98	3.31	3.40	3	4.00	0	3.92	1.91	1.81	86.9	62.6	73.7
53	2.93	3.34	2.87	3	4.67	0	4.80	2.18	1.69	90.8	66.3	67.3
54	2.89	2.96	3.32	3	3.67	0	2.27	1.48	1.09	90.8	55.1	58.2
55	2.85	2.43	3.60	2	4.50	1	3.70	2.42	1.79	73.5	59.2	55.1

56	2.75	2.85	2.67	3	6.17	0	2.63	1.69	1.76	83.7	64.3	65.3
57	2.67	3.19	2.34	2	4.33	1	3.70	2.42	1.79	62.2	60.2	67.3
58	2.60	3.27	2.31	3	4.00	0	2.02	1.32	0.98	85.9	53.5	41.4
59	2.53	2.36	3.35	3	4.67	0	1.91	1.25	0.92	82.7	45.9	38.8
60	2.50	2.69	3.08	3	4.33	0	1.91	1.25	0.92	84.7	40.8	35.7

Note. Seq Num: Sequence number; Ton All: mean overall tonality rating; Ton 1st: mean first-half tonality rating; Ton 2nd: mean second-half tonality rating; Cont Dir: number of contour direction changes; Ave Int: mean interval size in semitones; Semi Pres: a pitch one semitone away from the standard tone; KK Stan: fitness value of the standard; KKC Down: fitness value of the comparison tone one semitone lower than the standard; KKC Up: fitness value of the comparison tone one semitone higher than the standard; Same Resp: percentage of correct responses for trials with same comparison tones; Down Resp: percentage of correct responses for trials with different comparison tones one semitone lower than the standard; Up Resp: percentage of correct responses for trials with different comparison tones one semitone higher than the standard.

approach, the key of each sequence was determined automatically using the Krumhansl-Schmuckler (K-S) key-finding algorithm (Krumhansl, 1990), a method which entails correlating the pitch classes in each sequence with probe-tone rating profiles (Krumhansl & Kessler, 1982). Krumhansl and Kessler's key profiles consist of fitness values for all 12 chromatic pitches in major and minor modes. These values are derived from experiments in which listeners were asked to rate how well a probe tone "fit into or went with" a preceding tonal context. The resulting profiles showed that $\hat{1}$, the root of the tonic triad, has the highest fitness value, followed by $\hat{5}$, the dominant or fifth of the tonic triad, and $\hat{3}$, the third of the tonic triad. On the other end of the spectrum, nondiatonic tones have the lowest fitness values. For the purpose of assessing the most salient key in the 60 sequences used in the experiment, correlations were performed between a 12-dimensional vector representing the frequency count of the pitches in the sequences (here amounting to either 1s or 0s for each of the 12 possible chromatic pitches) and the Krumhansl-Kessler key profiles of all 24 major and minor keys. The key with the highest correlation value, whether major or minor, was deemed the designated key for the sequence. Although there are other key-finding algorithms that take order effects or memory decay into account, there was little reason to use a more complex algorithm—such as ones proposed by Temperley (2002) or Chew (2006)—given the simplicity and brevity of the melodies. Furthermore, from a qualitative perspective, the algorithm's output matched the authors' perceptions well. The only exceptions were sequences 24, 26, and 50, which were labeled as E major instead of A minor by the algorithm. This was likely due to the fact that minor-key profiles favor natural-minor scale degrees, and all three sequences included the raised $\hat{6}$ and $\hat{7}$ scale degrees for A minor. For the sake of consistency, the labels produced by the algorithm were retained for these sequences.

In the subsequent analysis, the subjective, whole-sequence tonality ratings were used as predictors of tonal strength instead of the objective K-S values (as defined above). The K-S values were used to assess the actual key and therefore the fitness of the standard and comparison tones to the sequences. The same K-S value could be used to assess tonal strength, but using the subjective ratings instead allows for a relatively independent second measure of tonal strength. Said another way, using the K-S to assess fitness and to assess tonal strength results in three unnecessarily redundant measures. The

subjective ratings provide an additional insight that is related to the K-S values. The correlation between the subjective ratings and the K-S values was $r(58) = .58$, $p < .001$, implying some but incomplete overlap ($r^2 = .34$).

Generalized linear mixed-effects models

The main analysis used GLMMs to assess various predictors of performance, with correct/incorrect responses as the binomial dependent variable for each trial (1 = correct, 0 = incorrect). Nine independent variables were initially considered as predictors: the three tonality ratings collected prior to the experiment (first half, second half, and whole sequence), the number of melodic contour direction changes in the sequence, the average interval size (defined as the mean of the absolute values of the semitone distances between intervals of the sequence), whether the interference tones contained a pitch that was one semitone (or octave equivalent) distant from the standard tone, the tonal fitness of the standard tone, the tonal fitness of the comparison tone, and the musical background of the listener. The contour changes, mean interval size, and semitone presence were included among the predictors because they were all factors implicated as influential in Deutsch's various pitch memory experiments. Figure 1 and Table 2 include all of the predictor values used in the subsequent analysis with the exception of musical experience, which does not apply to individual sequences.

The tonal fitness values of the standard and comparison tones were determined by first obtaining their Krumhansl-Kessler key profile values in the context of the designated key. For example, in the key of C major, C = 6.35; C# = 2.23; D = 3.48; D# = 2.33; E = 4.38; F = 4.09; F# = 2.52; G = 5.19; G# = 2.39; A = 3.66; A# = 2.29; B = 2.88. These values were then multiplied by the correlation coefficient associated with the salient key of the sequence. This was done to account for the fact that many of the sequences were not strongly tonal. In other words, the fitness values were scaled according to the strength of the perceived key (cf. Frankland & Cohen, 1996). For example, Sequence 1 prior to transposition (shown in Figure 1) contains all seven diatonic pitches in the C major scale resulting in the correlation values shown in Table 3. Thus for Sequence 1, the fitness value of the standard tone C is 4.80 (6.35 × .756), and the fitness values of the comparison tones B, C, and C# are 2.18 (2.88 × .756), 4.80 (6.35 × .756), and 1.69 (2.23 × .756) respectively. The correlation coefficient of .756 happened to be the highest for any given sequence (mean $r = .704$, $SD = .104$, min = .355 for all 60 sequences). The mean fitness values for standard tones and comparison tones across all conditions were 3.24 ($SD = .98$) and 2.42 ($SD = .99$) respectively.

The musical background metric was calculated from a formula that was the weighted sum of three values: years of training on a primary musical instrument, years of formal music theory training in college and graduate school, and self-rated overall musical experience. The instrumental and theory training values were converted to total decades (years divided by 10), and overall self-rating to a normalized value ranging from 0 to 1 (the 0–5 rating divided by 5). Instrument training and self-rated experience were given equal weight in the formula (multiplying by 0.3) and music theory training was given more weight (multiplying by 0.5). The rationale behind this system was to put more weight on formal academic music training over instrument lessons as a way of better differentiating among a pool of participants in which most had some musical training.

Table 3
Krumhansl-Kessler Key Profile Correlations for Sequence 1.

	Correlations	
Key	Major	Minor
C	.756	.092
G	.677	.116
D	.403	.523
A	.233	.712
E	.066	.589
B	-.291	.266
F#/Gb	-.755	-.197
C#/Db	-.716	-.333
G#/Ab	-.533	-.428
D#/Eb	-.379	-.700
A#/Bb	-.005	-.588
F	.546	-.053
Highest Correlation: C Major, .756		

The reason behind favoring academic training was to distinguish between musicians who were more consciously aware of tonal functions due to formal instruction from those who had a more implicit understanding of tonality (i.e., more similar to nonmusicians). The choice of weights was, by necessity, arbitrary since there was no empirically justified way to specify them. The general idea was to weight years of music theory experience significantly more, but to keep it to less than double the weight of performance experience. The resulting metric ranged from a minimum of 0 to a maximum of 1.5 (M = 0.67, SD = 0.33).

Evaluating individual predictors. In order to determine the relevant predictors, a full model with all predictors and interactions as fixed effects and subjects and sequences as random effects was first considered. A leave-one-out approach was taken, and predictors were retained if their exclusion resulted in a model with a higher Bayesian Information Criterion (BIC). The best resulting model (the one with the lowest BIC), retained four predictors: overall tonality rating, fitness of the standard, fitness of the comparison, and semitone presence. The complete results for this model (labeled Model 1) are shown in Table 4.

All of the predictors included were significant. The odds ratio for overall tonality indicated that for every one-point increase in tonality rating, the odds of responding correctly increased by a factor of 1.85. Increase in the fitness values of the standard and comparison tones also increased the odds of a correct answer by 1.13 for a one-point increase in the fitness of the standard tone and 1.30 for the fitness of the comparison tone. The last predictor, semitone presence, corresponded negatively to performance: the odds of a correct response decreased from 1 to 0.65 when an interference tone a semitone away from the standard was present. The negative effect aligns with the findings of earlier pitch memory studies by Deutsch.

Table 4a
Model 1 Results: Correct/Incorrect as DV.

	Var.	Std. Dev.	β	*SE* β	*t*	*p*	e^{β} (odds ratio)	95% CI
Model 1								
Random effects								
Sequence (Intercept)	0.07	0.27						
Subject (Intercept)	1.05	1.03						
Fixed effects								
Intercept			-1.53	0.34	-4.50	<.001		
Tonality overall			0.62	0.09	6.70	<.001	1.85	1.54-2.23
Fitness standard			0.12	0.05	2.64	.008	1.13	1.03-1.24
Fitness comparison			0.26	0.02	10.98	<.001	1.30	1.24-1.36
Semitone present = 1			-0.43	0.09	-4.98	<.001	0.65	0.55-0.77

Note. Number of observations = 17640; *c* statistic = 0.77; Somers's D_{xy} = 0.54.

Table 4b
Model 1 Comparison with Null Model.

	df	AIC	BIC	χ^2	*df*	*p*
Null model	3	16833	16856			
Model 1	7	16658	16713	183.00	4	<.001

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. The null model includes only the intercept and random effects.

Table 4c
Model 1 Observed and Predicted Frequencies.

	Observed		% Correctly
Predicted	Incorrect	Correct	Predicted
Incorrect	893	714	55.57%
Correct	3298	12735	79.43%
Overall % correct			77.26%

Note. Sensitivity = 12735/(12735+714)% = 94.69%. Specificity = 893/(893+3298)% = 21.31%. False positives = 3298/(3298+12735)% = 20.57%. False negatives = 714/(714+893)% = 44.43%.

Model evaluation. Multicollinearity of predictors was evaluated by determining the variance inflation factors (VIFs) for numeric predictors. Various recommendations for acceptable levels of VIF exist, the most common of which is 10; however, a conservative threshold is 3 (Zuur, Ieno, & Elphick, 2010). In this case, the VIFs were all 1.3 or lower, indicating that collinearity was not a concern.

The model was evaluated by comparing it to a null model consisting only of the intercept and the two random effects. The Akaike Information Criterion (AIC) and BIC values were both higher for the null model—indicating that the current model was an

improvement—and results of a chi-square test comparing the models were significant (Table 4b). The *c* static provided additional support for the validity of the model. The *c* statistic, also known as the area under the receiver operating characteristic (ROC) curve, is used as a measure of a model's discriminability. A *c* statistic of 0.5 means a model predicts the outcome no better than chance; a value of 1 means the model perfectly categorizes all responses. For the current model, the *c* statistic was .77, which indicated a good model. To further assess the predicted probabilities, a classification table using a cutoff point of .5 (to classify predicted probabilities as 0 or 1) is provided in Table 4c. The classification table shows that predictions of correct responses were more accurate than incorrect responses. This observation is supported by the difference between sensitivity (94.69%) and specificity (21.31%). The false positive rate (20.57%) was also less than the false negative rate (44.43%).

In summary, the model indicated that a stronger tonal context corresponded to better performance on the pitch recognition task. Higher fitness values for both the comparison and standard tones also increased the odds of a correct response. The presence of a pitch a semitone away from the standard, on the other hand, corresponded to worse performance.

Predicting response type. The next step was to model *response type* instead of correctness (Model 2). In other words, instead of trying to predict whether a response was correct or incorrect, the model predicted whether the response choice for a given trial was same or different (same = 1). In this case, there were twice as many different trials compared to same trials since there were two types of different comparison tones. The idea was to better understand which (if any) predictors might be influential in how a listener determined whether a comparison tone differed from a standard tone regardless of actual performance accuracy in the pitch memory task. The same procedure for determining predictors described above was used to choose the optimal model. The results are shown in Table 5. In addition to the predictors retained in the previous model, there were three additional ones: musical experience, an interaction between musical experience and overall tonality, and an interaction between and musical experience and fitness of the comparison tone.

Due to the addition of the interaction terms, multicollinearity was a potential issue. In order to address this, the predictors were centered (at zero mean) before the analysis was performed (Cohen, Cohen, West, & Aiken 2003; Gelman & Hill, 2007). After centering, all VIFs were low (maximum 1.2), indicating that collinearity was addressed. A *c* statistic of .81 showed that the predictive power of the model was strong.

All predictors were significant except musical experience. The odds ratio for overall tonality indicated that for every one-point increase in tonality rating, the odds of responding "same" decreased by a factor of 0.73. In other words, a better tonal structure enabled listeners to hear more clearly when a different comparison did not fit a given context. Similarly, a one-unit increase in the fitness of the standard decreased the odds of a "same" response by a factor of 0.65. In light of the bias toward "same" responses noted earlier, this indicates that a standard tone with a higher degree of fitness in a sequence was likely to be retained more accurately, enabling a better assessment. In contrast, the higher the fitness value of the comparison, the more likely the response was "same"; the odds of a "same" response increased by a factor of 2.31 for every unit increase in the fitness value of the comparison tone (with musical experience fixed at the mean).

The beta values for the two interactions represented the difference between the log-odds ratios corresponding to a unit increase in musical experience for tonality ratings differing by 1 and comparison fitness values differing by 1. In the case of the interaction between musical experience and comparison fitness, that difference translated to the odds of a "same" response increasing by a factor of 1.87. The opposite effect was found for the interaction between musical experience and tonality ratings; the odds of a "same" response was decreased by a factor of 0.68.

In summary, the results of this model showed that the bias toward "same" responses was related to the fitness of the comparison tone—not the fitness of the standard tone or the tonality of the sequence. The interaction between musical experience and comparison fitness indicated greater odds of a "same" response with more musical training and higher fitness values. However, the odds ratio of 1.87 is less than that of the comparison alone, 2.31, showing that musical training is actually mitigating the fitness effect. The classification table (Table 5c) also showed that there was little bias in the model. Both same and different trials were predicted with similar accuracy (74.57% and 73.06% respectively). Sensitivity (69.27%) was a little lower than specificity (77.92%), and false positive and false negatives rates were very close (25.43% and 26.94% respectively).

Although it is not possible to compare the same/different model to the previous correct/incorrect model directly, the models are complementary. They offer different insights into how the listeners responded. Model 1 essentially replicated prior work showing that a tonal context improves pitch memory performance. Model 2 revealed that the fitness value of the comparison tone had a strong influence in how listeners responded to the stimuli, offering a possible explanation for the "same" response bias.

Additional models. In order to better understand the relationship between performance and response type, two more models using different sets of trials were utilized (labeled Models 3 and 4). Model 3 only included trials with "same" responses (N = 8523) and Model 4 only included trials with "different" responses (N = 9117); the dependent variables for both were correct/incorrect. The results are shown in Tables 6 and 7. Collinearity was not a concern (all VIFs <= 1.6), and the data did not need to be normalized. The two models fit the data very well, with *c* statics of .91 for both.

Model 3 retained the same predictors as Model 1: overall tonality, fitness of the standard and comparison tones, and semitone presence. All effects were significant except overall tonality. The odds ratio for overall tonality indicated that for every one-point increase in tonality rating, the odds of responding correctly increased by a factor of 1.11. Increases in the fitness values of the standard and comparison tones corresponded to very different outcomes: the odds of a correct answer were decreased by a factor of 0.35 for a one-point increase in the fitness of the standard, while they increased very substantially—by a factor 10.62—for every one-point increase in the fitness of the comparison. Semitone presence, again, corresponded negatively to performance, with an odds ratio of 0.30.

Model 4 included an extra predictor, musical experience, in addition to the Model 1 predictors. The odds of a correct response increased by a factor of 2.38 for a one-point increase in tonality rating. A unit increase in the fitness of the standard tone increased the odds of a correct response by a factor of 1.67. In contrast, the odds of correct response decreased dramatically—by a factor of .08—for a unit increase in comparison-tone

Table 5a

Model 2 Results: Response Type (Same/Different) as DV.

	Var.	Std. Dev.	β	*SE* β	*t*	*p*	e^{β} (odds ratio)	95% CI
Model 2								
Random effects								
Sequence (Intercept)	0.13	0.35						
Subject (Intercept)	0.34	0.58						
Fixed effects								
Intercept			0.16	0.11	1.50	0.145	2.61	1.00-7.02
Tonality overall			-0.57	0.11	-5.10	<.001	0.73	0.56-0.96
Fitness standard			-0.43	0.06	-7.70	<.001	0.65	0.58-0.73
Fitness comparison			1.26	0.02	51.60	<.001	2.31	2.11-2.54
Semitone present = 1			-0.50	0.11	-4.70	<.001	0.61	0.49-0.75
Musical experience			-0.49	0.26	-1.90	0.064	0.52	0.20-1.36
Mus. ex.*Tonality			-0.38	0.12	-3.20	0.002	0.68	0.54-0.86
Mus. ex.*Fit. cmp.			0.63	0.07	9.40	<.001	1.87	1.64-2.13

Note. Number of observations = 17640; *c* statistic = 0.81; Somers's D_{xy} = 0.61. Mus. ex. = musical experience; Fit. cmp. = Fitness of comparison tone.

Table 5b

Model 2 Comparison with Null Model.

	df	AIC	BIC	χ^2	*df*	*p*
Null model	3	23469	23493			
Model 2	10	19717	19795	3766	7	<.001

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. The null model includes only the intercept and random effects.

Table 5c

Model 2 Observed and Predicted Frequencies.

	Observed		% Correctly
Predicted	Different	Same	Predicted
Different	7104	2619	73.06%
Same	2013	5904	74.57%
Overall % correct			73.74%

Note. Sensitivity = 5904/(5904+2619)% = 69.27%. Specificity = 7104/(7104+2013)% = 77.92%. False positives = 2013/(2013+5904)% = 25.43%. False negatives = 2619/(2619+7104)% = 26.94%.

Table 6a

Model 3 Results: Correct/Incorrect as DV and Only Trials with "Same" Responses Included.

	Var.	Std. Dev.	β	*SE* β	*t*	*p*	e^{β} (odds ratio)	95% CI
Model 3								
Random effects								
Sequence (Intercept)	0.15	0.39						
Subject (Intercept)	0.67	0.82						
Fixed effects								
Intercept			-2.14	0.50	-4.30	<.001		
Tonality overall			0.11	0.15	0.70	0.470	1.11	0.84-1.49
Fitness standard			-1.05	0.08	-13.00	<.001	0.35	0.30-0.41
Fitness comparison			2.36	0.06	42.70	<.001	10.62	9.55-11.86
Semitone present = 1			-1.21	0.14	-9.00	<.001	0.30	0.23-0.39

Note. Number of observations = 8523; *c* statistic = 0.91; Somers's D_{xy} = 0.82.

Table 6b

Model 3 Comparison with Null Model.

	df	AIC	BIC	χ^2	*df*	*p*
Null model	3	10289	10310			
Model 3	7	6728	6777	3569.00	4	<.001

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. The null model includes only the intercept and random effects.

Table 6c

Model 3 Observed and Predicted Frequencies.

	Observed		% Correctly
Predicted	Incorrect	Correct	Predicted
Incorrect	2804	405	87.38%
Correct	609	4705	88.54%
Overall % correct			88.10%

Note. Sensitivity = 4705/(4705+405)% = 92.07%. Specificity = 2804/(2804+609)% = 82.16%. False positives = 609/(609+4705)% = 11.46%. False negatives = 405/(405+2804)% = 12.62%.

Table 7a

Model 4 Results: Correct/Incorrect as DV and Only Trials with "Different" Responses Included.

	Var.	Std. Dev.	β	*SE* β	*t*	*p*	e^{β} (odds ratio)	95% CI
Model 4								
Random effects								
Sequence (Intercept)	0.84	0.92						
Subject (Intercept)	0.80	0.90						
Fixed effects								
Intercept			2.49	1.06	2.36	0.018	12.06	1.51-102.34
Tonality overall			0.87	0.30	2.86	0.004	2.38	1.29-4.36
Fitness standard			0.52	0.15	3.36	<.001	1.67	1.24-2.29
Fitness comparison			-2.48	0.09	-27.97	<.001	0.08	0.07-0.10
Semitone present = 1			1.79	0.30	6.01	<.001	5.99	3.33-10.91
Musical experience			1.36	0.43	3.14	0.002	3.88	1.64-9.38

Note. Number of observations = 9117; *c* statistic = 0.91; Somers's D_{xy} = 0.82.

Table 7b

Model 4 Comparison with Null Model.

	df	AIC	BIC	χ^2	*df*	*p*
Null model	3	4780	4802			
Model 4	8	3196	3253	1594	5	<.001

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. The null model includes only the intercept and random effects.

Table 7c

Model 4 Observed and Predicted Frequencies.

	Observed		% Correctly
Predicted	Incorrect	Correct	Predicted
Incorrect	386	81	82.66%
Correct	392	8258	95.47%
Overall % correct			94.81%

Note. Sensitivity = 8258/(8258+81)% = 99.03%. Specificity = 386/(386+392)% = 49.61%. False positives = 392/(392+8258)% = 4.53%. False negatives = 81/(81+386)% = 17.34%.

fitness. Surprisingly, semitone presence had a positive relationship to performance, with the odds of a correct response increasing by a factor of 5.99. Increase in musical experience also corresponded to correct responses, although the seemingly substantial odds ratio of 3.88 is somewhat misleading given that the metric had a maximum value of 1.5 (in other words, a one-point increase spans a significant portion of the range).

In summary, the results of Models 3 and 4 highlighted the factors influencing "same" versus "different" responses. For trials with "same" responses, the fitness of the comparison significantly promoted a correct (= "same") response. The tonality of the sequence and fitness of the standard, on the other hand, had little (or negative) impact. For trials with "different" responses, the fitness of the standard and the overall tonality promoted a correct (= "different) response, but the fitness of the comparison resulted in worse performance.

Atonal contexts. The models discussed thus far all point to the fitness of the comparison tone being the primary predictor of listener responses. However, in weaker tonal contexts, the fitness values overall should be weaker by definition (cf. Vuvan et al., 2014). The final model (Model 5, $N = 2646$) included only sequences with mean overall tonality ratings of less than 3.0—i.e., the nine lowest-rated sequences. As in the case of Models 1, 3, and 4, the dependent variable was correct/incorrect (see Table 8 for results). The model included comparison fitness, semitone presence, and musical experience as predictors, and one interaction effect between comparison fitness and semitone presence. Due to the interaction, the variables were centered to address multicollinearity, resulting in VIFs that were sufficiently low (maximum 2.3).

All of the predictors were significant, although closer inspection of the stimuli suggests that the interaction effect was likely coincidental: there are only two sequences in the semitone-presence category, both of which have coincidentally higher comparison-fitness values compared to the other seven sequences. The main effects of semitone presence and musical experience were not particularly surprising. The odds of a correct response decreased by a factor of 0.56 for the former (again, supporting Deutsch's findings), and the odds of a correct response increased by a factor 2.38 for every unit increase in musical experience. However, the strong effect of comparison fitness once again—even in very weak (or nonexistent) tonal contexts—was surprising: the odds of a correct response increased by a factor of 3.03 for a unit increase in comparison fitness. These results suggested that even in non-tonal contexts, listeners were unconsciously applying tonal templates as a way of encoding more efficient representations of melodic sequences.

Discussion

The objective of this study was to reexamine the effect of tonality on short-term pitch memory in order to better understand the contexts in which tonality has a negative versus positive impact on pitch recognition. Tonality ratings collected for 60 short melodies were used as predictors for a subsequent pitch-memory experiment that featured the melodies with comparison tones appended. The data were analyzed using generalized linear mixed-effects models (GLMMs) to model listener responses given predictors chosen from nine variables, including the tonality ratings and the tonal fitness values of the comparison and standard tones. When the dependent variable was correct/incorrect

Table 8a

Model 5 Results: Atonal Trials Only (Correct/Incorrect as DV).

	Var.	Std. Dev.	β	*SE* β	*t*	*p*	e^{β} (odds ratio)	95% CI
Model 5								
Random effects								
Sequence (Intercept)	0.49	0.70						
Subject (Intercept)	0.01	0.08						
Fixed effects								
Intercept			1.09	0.13	8.72	<.001	2.97	2.32-3.83
Fitness comparison			1.11	0.10	11.44	<.001	3.03	2.53-3.71
Musical experience			0.87	0.34	2.59	0.010	2.38	1.22-4.68
Semitone present = 1			-0.58	0.14	-4.10	<.001	0.56	0.41-0.76
Fitness cmp*Semitn			-0.92	0.15	-6.17	<.001	0.40	0.30-0.53

Note. Number of observations = 2646; *c* statistic = 0.77; Somers's D_{xy} = 0.53. Fitness cmp*Semitn = interaction between fitness of comparison tone and presence of semitone.

Table 8b

Model 5 Comparison with Null Model.

	df	AIC	BIC	χ^2	*df*	*p*
Null model	3	3244	3262			
Model 5	7	3054	3096	198.00	4	<.001

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. The null model includes only the intercept and random effects.

Table 8c

Model 5 Observed and Predicted Frequencies.

	Observed		% Correctly
Predicted	Incorrect	Correct	Predicted
Incorrect	406	204	66.56%
Correct	519	1517	74.51%
Overall % correct			72.68%

Note. Sensitivity = 1517/(1517+204)% = 88.15%. Specificity = 406/(406+519)% = 43.89%. False positives = 519/(519+1517)% = 25.49 %. False negatives = 204/(204+406)% = 33.44%.

(Model 1), strength of tonality increased the odds of a correct response (from 1 to 1.85 for a one-unit increase in tonality rating). Higher fitness values for both the comparison and standard tones also increased the odds of a correct response (an odds ratio of 1.13 for the standard and 1.30 for the comparison).

These results on the outset seemed to support the theory that a stronger tonal context yields better performance as well as contradict the notion that fitness of the comparison has a negative effect on accurate recognition. However, the results were not particularly surprising since trials with in-key, different comparisons encompassed only a subset of the stimuli. A strong tonal context had the opposite effect when there were out-of-key comparison tones, in which case the low degree of fitness of the out-of-key comparisons appeared to aid listeners in detecting mismatches.

Model 1 essentially mimicked Frankland and Cohen's (1996) approach using different techniques. Frankland and Cohen explored how the strength of a tonal context and the role of standard and comparison tones in that context might predict pitch recognition performance. As in the current work, they used the K-S key-finding algorithm to find the best-fitting key for each melodic sequence. They then determined the stability of the standard tone and expectancy of the comparison tone given the designated key, and used those values to predict performance. "Stability" and "expectancy" are analogous to "fitness" here. Similar to the current Model 1 results, Frankland and Cohen concluded that their standard and comparison tone models also predicted performance well, though neither was markedly more effective than the other.

Model 2 (response type same/different as the dependent variable) provided new evidence for the hypothesis that fitness of the comparison tone was an important factor in how listeners responded to the pitch memory task. The results showed that all of the predictors except fitness of the comparison and the interaction between comparison fitness and musical experience decreased the odds of a "same" response. For every unit increase in comparison fitness, the odds of a "same" response increased substantially (by a factor of 2.31). In combination, Models 1 and 2 presented complementary perspectives: Model 1 replicated prior work showing that a tonal context improves pitch memory performance and Model 2 showed that the fitness of the comparison tone was a crucial factor in listener responses.

Models 3 and 4 provided additional evidence illuminating how comparison-tone fitness affected pitch memory performance. These models included only trials with "same" and "different" responses respectively, with incorrect/correct as the dependent variable. Again, the fitness of the comparison tone was the most notable factor. The two models provided starkly contrasting results, with higher fitness values corresponding better performance in "same" trials (odds ratio 10.62) and dramatically worse performance in "different" trials (odds ratio .08).

One additional model, Model 5, included only sequences with the lowest tonality ratings. Surprisingly, the fitness of the comparison was again a significant predictor of performance. It is possible that despite the low tonality ratings, the sequences were short enough for listeners to construe some type of vague tonal context. It is also arguable that the tonality ratings were not low enough to be truly atonal since none of the mean ratings were lower than 2.5 (on a scale of 1-5). Regardless, these results further reinforce the conclusion that the fitness of the comparison was a powerful influence on how listeners responded to the stimuli.

Fitness values are inherently tied to perceived tonality; there is no way to completely divorce the two percepts because the concept of fitness itself assumes the perception of a tonal context. While other possible fitness measures are possible, such as how a tone fits within the expected contour of a melody either using Gestalt principles or other theories of melodic expectation (Narmour, 1990, 1992), that is beyond the scope of this work. Furthermore, most models of melodic expectation (e.g., Bharucha, 1996, Larson, 2004; Margulis, 2005; Pearce & Wiggins, 2006) take tonality into account in some form. However, as a follow-up, we did explore a model that added intervallic distance between the final note of the sequence and the comparison tone (in semitones) as a predictor. This additional predictor had a corresponding odds ratio of 1.03 and was not significant ($p = .07$).

While the memory of individual pitches cannot be equated with the memory of sequences of pitches (the latter of which may also be transposed), there are inherent similarities between the two types of recognition tasks in a tonal context. One observation from Cuddy et al. (1979)'s melodic memory study parallels the results discussed here. Cuddy et al. examined the recognition of transposed, three-note sequences in different melodic contexts. The transposed comparison melodies were either identical to the original melodies in intervallic contour or altered such that one pitch was one semitone off. The three-note sequences were either embedded within a larger melodic context (two notes preceding and following the core three notes) or not. The contextual melodic material was designed to be diatonic ending on a tonic, diatonic not ending on a tonic, or non-diatonic, defined as not having all seven pitches belonging to a single diatonic scale. In agreement with other studies, the results indicated that the more tonal-sounding the context, the easier it was for listeners to spot pitch changes. They noted (in Experiment 3) one "anomalous" sequence in the most tonal category where the altered tone was raised by a semitone (within-key) from the mediant to the subdominant ($\hat{3}$ to $\hat{4}$). This particular stimulus had a very high error rate, resulting in performance equivalent to chance. This is conceptually analogous to the types of errors we observed in the present study. The difference is that Cuddy et al. tested recognition of transposed melodies, while individual, untransposed pitches were tested in the current study.

This mutability of pitch encoding in a tonal context is also reflected in an experiment by Krumhansl (1979, Experiment 1) that did not employ a recognition task at all: listeners heard two tones and were asked to rate how similar the first one was to the second when presented in a strong tonal context. The tonal context was established by playing a C major scale or triad prior to the presentation of the tones. The results showed an asymmetric relationship between the two pitches; nondiatonic tones were judged to be more similar to diatonic tones when the nondiatonic tone came first. Reflecting on these results, Krumhansl (1983) hypothesized that the memory representation of nondiatonic tones were "unstable, tending to become assimilated over time into more stable elements within the tonal system." This tendency toward assimilation in an established tonal context is reflected in the results presented here. An unstable standard tone (e.g., a leading tone) coupled with a stable but different comparison tones (e.g., tonic) resulted in poor recognition performance, likely due in part to the assimilation of the unstable tone into the tonal context. Subsequently, the deceptive stability of the comparison tone in that context would further complicate the differentiation of the two pitches.

In summary, there are two main conclusions that can be drawn from the current

results: (1) The fitness of the comparison—that is, how well the tone fit in a tonal context—was a key factor in how listeners responded in the pitch memory task. Comparison tones with higher fitness values increased performance when the comparison and standard were the same, but decreased performance when they were different. (2) The presence of an interference tone one semitone away from the standard corresponded to a decrease in performance. The first conclusion reflects what is likely the primary strategy used by listeners when determining a response in a pitch memory task. The second conclusion highlights the only non-tonal factor derived directly from the sequences that significantly predicted performance. As a whole, these results provide conclusive evidence for how a tonal context can facilitate or impede pitch memory.

References

Bigand, E. (1990). Abstraction of two forms of underlying structure in a tonal melody. *Psychology of Music*, *18*(1), 45–59. http://doi.org/10.1177/0305735690181004

Boltz, M. (1989). Perceiving the end: Effects of tonal relationships on melodic completion. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 749–761. http://doi.org/10.1037/0096-1523.15.4.749

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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Chew, E. (2006). Slicing it all ways: Mathematical models for tonal induction, approximation, and segmentation using the Spiral Array. *INFORMS Journal on Computing*, *18*(3), 305–320.

Cohen, A. J., Thorpe, L. A., & Trehub, S. E. (1987). Infants' perception of musical relations in short transposed tone sequences. *Canadian Journal of Psychology*, *41*(1), 33–47.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Croonen, W. L. M. (1994). Effects of length, tonal structure, and contour in the recognition of tone series. *Perception & Psychophysics*, *55*(6), 623–632. http://doi.org/10.3758/BF03211677

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

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

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

Deutsch, D. (1972a). Effect of repetition of standard and comparison tones on recognition memory for pitch. *Journal of Experimental Psychology*, *93*(1), 156–162.

Deutsch, D. (1972b). Mapping of interactions in the pitch memory store. *Science*, *175*, 1020–1022.

Deutsch, D. (1973a). Interference in memory between tones adjacent in the musical scale. *Journal of Experimental Psychology*, *100*(2), 228–231.

Deutsch, D. (1973b). Octave generalization of specific interference effects in memory for tonal pitch. *Perception & Psychophysics*, *13*(2), 271–275.

Deutsch, D. (1974). Generality of interference by tonal stimuli in recognition memory for pitch. *Quarterly Journal of Experimental Psychology*.

Deutsch, D. (1978). Delayed pitch comparisons and the principle of proximity. *Perception & Psychophysics*, *23*(3), 227–230. http://doi.org/10.3758/BF03204130

Dewar, K. M., Cuddy, L. L., & Mewhort, D. J. K. (1977). Recognition memory for single tones with and without context. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(1), 60–67.

Dibben, N. (1994). The cognitive reality of hierarchic structure in tonal and atonal music.

Music Perception, *12*(1), 1–25.
Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*, 341–354.
Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception & Psychophysics*, *50*(4), 305–313. http://doi.org/10.3758/BF03212222
Dowling, W. J., Kwak, S., & Andrews, M. W. (1995). The time course of recognition of novel melodies. *Perception & Psychophysics, 57*(2), 136–149.
Francès, R. (1988). *The perception of music*. (W. J. Dowling, Trans.). Hillsdale, NJ: Erlbaum.
Frankland, B. W., & Cohen, A. J. (1996). Using the Krumhansl and Schmuckler key-finding algorithm to quantify the effects of tonality in the interpolated-tone pitch-comparison task. *Music Perception*, *14*(1), 57–83.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.
Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36,* ECVP Abstract Supplement.
Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, *11*, 346–374.
Krumhansl, C. L. (1983). Perceptual structures for tonal music. *Music Perception*, *1*(1), 28–62.
Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. New York: Oxford University Press.
Larson, S. (2004). Musical forces and melodic expectations: Comparing computer models and experimental results. *Music Perception*, *21*(4), 457–498.
Long, P. A. (1977). Relationships between pitch memory in short melodies and selected factors. *Journal of Research in Music Education*, *25*(4), 272–282. http://doi.org/10.2307/3345268
Margulis, E. H. (2005). A model of melodic expectation. *Music Perception*, *22*(4), 663–714.
Narmour, E. (1990). *The analysis and cognition of basic melodic structures*. Chicago: University of Chicago Press.
Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realization model*. Chicago: University of Chicago Press.
Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*(5), 377–405. http://doi.org/10.1525/mp.2006.23.5.377
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.
Serafine, M. L., Glassman, N., & Overbeeke, C. (1989). The cognitive reality of hierarchic structure in music. *Music Perception*, 397–430.
Taylor, J. A. (1976). Perception of tonality in short melodies. *Journal of Research in Music Education*, *24*(4), 197–208. http://doi.org/10.2307/3345130
Temperley, D. (2002). A Bayesian approach to key-finding. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and Artificial Intelligence. Lectures in Computer Science* (Vol. 2445, pp. 195–206). Berlin.

Trainor, L. J., & Trehub, S. E. (1992). A comparison of infants' and adults' sensitivity to Western musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(2), 394–402. http://doi.org/10.1037/0096-1523.18.2.394

Trainor, L. J., & Trehub, S. E. (1993). Musical context effects in infants and adults: Key distance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(3), 615–626. http://doi.org/10.1037/0096-1523.19.3.615

Trehub, S. E., Cohen, A. J., Thorpe, L. A., & Morrongiello, B. A. (1986). Development of the perception of musical relations: Semitone and diatonic structure. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(3), 295–301. http://doi.org/10.1037/0096-1523.12.3.295

Trehub, S. E., Schellenberg, E. G., & Kamenetsky, S. B. (1999). Infants' and adults' perception of scale structure. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 965–975. http://doi.org/10.1037/0096-1523.25.4.965

Vuvan, D. T., Podolak, O. M., & Schmuckler, M. A. (2014). Memory for musical tones: the impact of tonality and the creation of false memories. *Frontiers in Psychology*, *5*, Article 582. http://doi.org/10.3389/fpsyg.2014.00582

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

Wickelgren, W. A. (1966). Consolidation and retroactive interference in short-term recognition memory for pitch. *Journal of Experimental Psychology*, *72*(2), 250–259. http://doi.org/10.1037/h0023438

Wickelgren, W. A. (1969). Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology*, *6*(1), 13–61. http://doi.org/10.1016/0022-2496(69)90028-5

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3-14.