# Pitch-Class Distribution and the Identification of Key

David Temperley and Elizabeth West Marvin
*Eastman School of Music of the University of Rochester*

**This study examines the *distributional* view of** key-finding, which holds that listeners identify key by monitoring the distribution of pitch-classes in a piece and comparing this to an ideal distribution for each key. In our experiment, participants judged the key of melodies generated randomly from pitch-class distributions characteristic of tonal music. Slightly more than half of listeners' judgments matched the generating keys, on both the untimed and the timed conditions. While this performance is much better than chance, it also indicates that the distributional view is far from a complete explanation of human key identification. No difference was found between participants with regard to absolute pitch ability, either in the speed or accuracy of their key judgments. Several key-finding models were tested on the melodies to see which yielded the best match to participants' responses.

*Received October 4, 2006, accepted September 4, 2007.*

**Key words:** key, key perception, probabilistic models, absolute pitch, music psychology

**How do listeners identify the key of a piece** as they hear it? This is surely one of the most important questions in the field of music perception. In tonal music, the key of a piece governs our interpretation of pitches and chords; our understanding of a note and its relations with other notes will be very different depending on whether it is interpreted as the tonic note (scale degree 1), the leading-tone (scale degree 7), or some other scale degree. Experimental work has shown that listeners' perception of key affects other aspects of musical processing and experience as well. Key context affects the memory and recognition of melodies (Cuddy, Cohen, & Mewhort, 1981; Cuddy, Cohen, & Miller, 1979; Marvin, 1997), conditions our expectations for future events (Cuddy & Lunney, 1995; Schmuckler, 1989), and affects the speed and accuracy with which notes can be processed (Bharucha & Stoeckig, 1986; Janata & Reisberg, 1988). For all of these reasons, the means whereby listeners identify the key of a piece is an issue of great interest.

Several ideas have been proposed to explain how listeners might identify key. One especially influential view of key-finding is what might be called the *distributional* view. According to this view, the perception of key depends on the distribution of pitch-classes in the piece. Listeners possess a cognitive template that represents the ideal pitch-class distribution for each major and minor key; they compare these templates with the actual pitch-class distribution in the piece and choose the key whose ideal distribution best matches that of the piece. While this idea has had numerous advocates, the distributional approach to key perception has had many critics as well. Some musicians and music theorists (in our experience) find the distributional view implausible, because it seems so unmusical and "statistical," and ignores all kinds of musical knowledge that we know to be important—knowledge about conventional melodic patterns, cadential gestures, implied harmonies, large-scale melodic shape, and so on. Critics of the distributional approach have argued that key perception depends crucially on pitch ordering and on the intervallic and scale-degree patterns that pitches form. We might call this general view of key-finding the *structural* view, as it claims a role for musical structure in key perception beyond the mere distribution of pitch-classes.

How can we test whether listeners use a distributional approach or a structural approach to key identification? In real music, both distributional and structural cues are present: the key may be identifiable by distributional means, but no doubt there are also structural cues that could be used to determine the key. Thus real music can tell us little about which strategy listeners are using. To answer this question, we would need to test listeners' key perceptions in musical stimuli designed to match the pitch-class distributions of each key but without any structural cues, or conversely, in stimuli that feature structural cues suggestive of a particular key but lacking the appropriate pitch-class distribution

*Music Perception* volume 25, issue 3, pp. 193–212, issn 0730-7829, electronic issn 1533-8312  DOI:10.1525/mp.2008.25.3.193

for that key. In the current study, we take the former approach: we examine listeners' perception of key in melodies generated randomly from pitch-class distributions drawn from a classical music corpus. Since the keys of such melodies are (presumably) not reliably indicated by structural cues, a high rate of success in key identification will suggest that listeners are using a distributional approach.

## Previous Studies of Key Identification

The modeling of key identification has been an active area of research for several decades. Perhaps the first attempt in this area was the monophonic key-finding model of Longuet-Higgins and Steedman (1971). Longuet-Higgins and Steedman's model processes a melody in a left-to-right fashion; at each note, it eliminates all keys whose scales do not contain that note. When only one key remains, that is the chosen key. If the model gets to the end of the melody with more than one key remaining, it chooses the one whose tonic is the first note of the melody, or failing that, the one whose dominant is the first note. If at any point all keys have been eliminated, the "first-note" rule again applies. In a test using the 48 fugue subjects of Bach's *Well-Tempered Clavier*, the model identified the correct key in every case. However, it is not difficult to find cases where the model would encounter problems. In "The Star-Spangled Banner," for example (Figure 1a), the first phrase strongly implies a key of Bb major, but the model would be undecided between Bb major, F major, and several other keys in terms of scales; invoking the first-note rule would yield an incorrect choice of F major. Another problem for the model concerns chromatic notes (notes outside the scale); the traditional melody "Ta-ra-ra-boom-de-ay" (Figure 1b) clearly conveys a tonal center of C, but the presence of the chromatic F# and D# would cause the model to eliminate this key. These examples show that key identification, even in simple tonal melodies, is by no means a trivial problem.

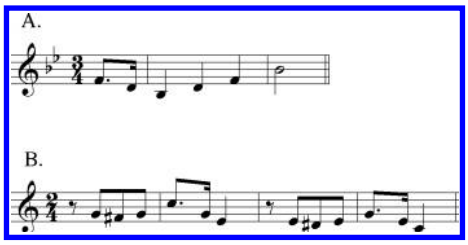


FIGURE 1. (A) "The Star-Spangled Banner." (B) "Ta-ra-ra-boom-de-ay."

An alternative approach to key-finding is a procedure proposed by Carol Krumhansl and Mark Schmuckler, widely known as the Krumhansl-Schmuckler (hereafter K-S) key-finding algorithm and described most fully in Krumhansl (1990). The algorithm is based on a set of "key-profiles," first proposed by Krumhansl and Kessler (1982), representing the stability or compatibility of each pitch-class relative to each key. The key-profiles are based on experiments in which participants were played a key-establishing musical context such as a cadence or scale, followed by a probe-tone, and were asked to judge how well the probe-tone "fit" given the context (on a scale of 1 to 7, with higher ratings representing better fitness). Krumhansl and Kessler averaged the rating across different contexts and keys to create a single major key-profile and minor key-profile, shown in Figure 2 (we will refer to these as the K-K profiles). The K-K key-profiles reflect some well accepted principles of Western tonality, such as the structural primacy of the tonic triad and of diatonic pitches over their chromatic embellishments. In both the major and minor profiles, the tonic pitch is rated most highly, followed by other notes of the tonic triad, followed by other notes of the scale (assuming the natural minor scale in minor), followed by chromatic notes.

Given these key-profiles, the K-S algorithm judges the key of a piece by generating an "input vector"; this is, again, a twelve-valued vector, showing the total duration of each pitch-class in the piece. The correlation is then calculated between each key-profile vector and the input vector; the key whose profile yields the highest correlation value is the preferred key. The use of correlation means that a key will score higher if the peaks of its key-profile (such as the tonic-triad notes) have high values in the input vector. In other words, the listener's sense of the fit between a pitch-class and a key (as reflected in the key-profiles) is assumed to be highly correlated with the frequency and duration of that pitch-class in pieces in that key.

The K-S model has had great influence in the field of key-finding research. One question left open by the model is how to handle modulation: the model can output a key judgment for any segment of music it is given, but how is it to detect changes in key? Krumhansl herself (1990) proposed a simple variant of the model for this purpose, which outputs key judgments for each measure of a piece, based on the algorithm's judgment for that measure (using the basic K-S algorithm) combined with lower-weighted judgments for the previous and following measures. Other ways of incorporating modulation into the K-S model have also been proposed (Huron & Parncutt, 1993; Schmuckler &

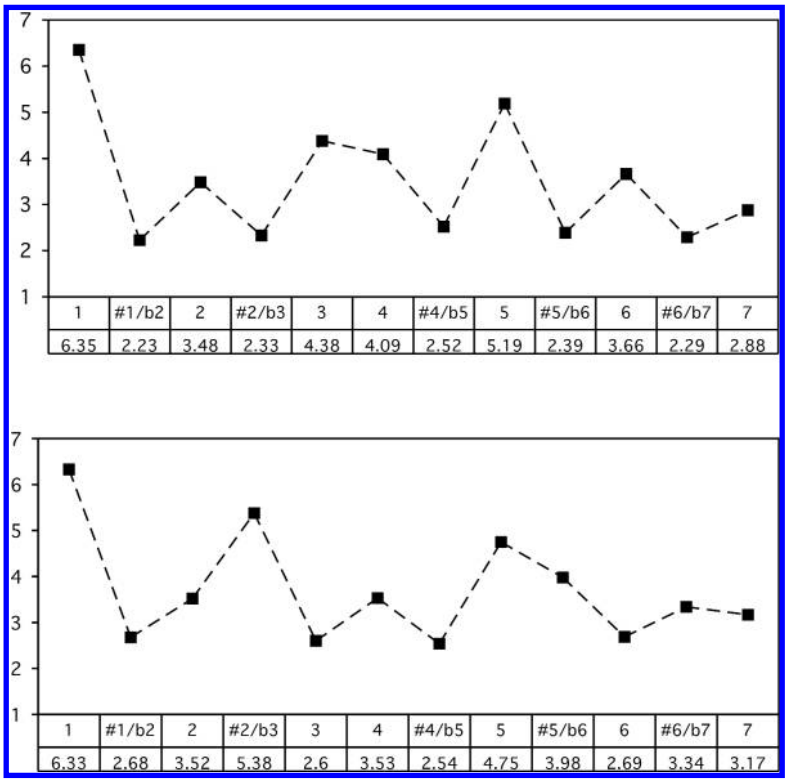


FIGURE 2. Key-profiles for major keys (above) and minor keys (below). From Krumhansl and Kessler (1982).

Tomovski, 2005; Shmulevich & Yli-Harja, 2000; Temperley, 2001; Toiviainen & Krumhansl, 2003). Other authors have presented models that differ from the K-S model in certain respects, but are still essentially distributional, in that they are affected only by the distribution of pitch-classes and not by the arrangement of notes in time. In Chew's (2002) model, pitches are located in a three-dimensional space; every key is given a characteristic point in this space, and the key of a passage of music can then be identified by finding the average position of all events in the space and choosing the key whose "key point" is closest. In Vos and Van Geenen's (1996) model, each pitch in a melody contributes points to each key whose scale contains the pitch or whose I, IV, or V7 chords contain it, and the highest scoring key is the one chosen. Yoshino and Abe's (2005) model is similar to Vos and Van Geenen's, in that pitches contribute points to keys depending on their function within the key; temporal ordering is not considered, except to distinguish "ornamental" chromatic tones from other chromatic tones. Finally, Leman's (1995) model derives key directly from an acoustic signal, rather than from a representation where notes have already been identified. The model is essentially a key-profile model, but in this case the input vector represents the strength of each pitch-class (and its harmonics) in the auditory signal; key-profiles are generated in a similar fashion, based on the frequency content of the primary chords of each key.

Temperley (2007) proposes a distributional key-finding model based on probabilistic reasoning. This probabilistic model assumes a generative model in which melodies are generated from keys. A key-profile in this case represents a probability function, indicating the probability of each scale-degree given a key. Such key-profiles can be generated from musical corpora; the profiles in Figure 3 are drawn from the openings of Mozart and Haydn string quartet movements (these profiles are discussed further below). Given such key-profiles, a melody can be constructed as a series of notes generated from the key-profile. The probability of the melody given a key, $P(\text{melody} \mid \text{key})$, is then the product of all the probabilities (key-profile values) for the individual notes. For example, given the key of C major, the

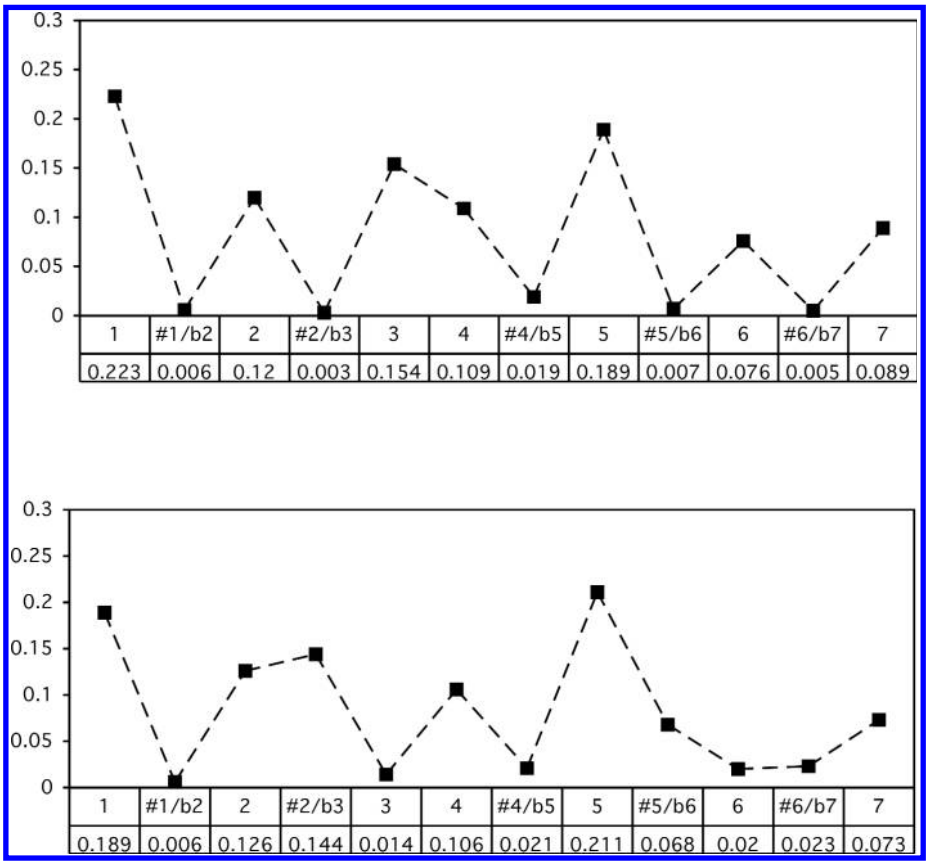


FIGURE 3. Key-profiles generated from the string quartets of Mozart and Haydn, for major keys (above) and minor keys (below).

probability for the melody C-F#-G (scale degrees 1-#4-5) would be $.223 \times .019 \times .189 = .00080$.

A basic rule of probability, Bayes' rule, then allows us determine the probability of any key given the melody, $P(\text{key} \mid \text{melody})$:

$$P(\text{key} \mid \text{melody}) = \frac{P(\text{melody} \mid \text{key})\, P(\text{key})}{P(\text{melody})} \qquad (1)$$

The denominator of the expression on the right, $P(\text{melody})$, is just the overall probability of a melody and is the same for all keys. As for the numerator, $P(\text{key})$ is the "prior" probability of each key occurring. If we assume that all keys are equal in prior probability, then this, too, is constant for all keys (we discuss this assumption further below). Thus

$$P(\text{key} \mid \text{melody}) \propto P(\text{melody} \mid \text{key}) \qquad (2)$$

To identify the most probable key given a melody, then, we simply need to calculate $P(\text{melody} \mid \text{key})$ for all 24 keys and choose the key yielding the highest value. This model was tested on a corpus of European folk songs, and identified the correct key in 57 out of 65 melodies.[1]

[1]The model described here is a somewhat simplified version of the monophonic key-finding model described in Chapter 4 of Temperley (2007). The model generates monophonic pitch sequences using factors of key, range, and pitch proximity, and can be used to model key-finding, expectation, and other phenomena. The model used here does not consider range and pitch proximity, but these factors have little effect on the model's key-finding behavior in any case (see Temperley, 2007, Chapter 4, especially Note 6). As Temperley notes (pp. 79-81), this approach to key-finding is likely to be less effective for polyphonic music; treating each note as generated independently from the key-profile is undesirable in that case given the frequent use of doubled and repeated pitch-classes. For polyphonic music, Temperley proposes instead to divide the piece into short segments and label each pitch-class as "present" or "absent" within the segment. For melodies, however, the approach of counting each note seems to work well.

Despite the number of researchers who have embraced the distributional approach to key-finding, not all have accepted it. Some have suggested that distributional methods neglect the effect of the temporal ordering of pitches in key perception. Butler and colleagues (Butler, 1989; Brown, Butler, & Jones, 1994) have argued that key detection may depend on certain "goal-oriented harmonic progressions" that are characteristic of tonal music. Butler et al. focus especially on tritones—what they call a "rare interval"—becauses tritones occur only between two scale degrees (4 and 7) within the major scale, whereas other intervals occur more often, between multiple scale degrees (e.g., an ascending perfect fourth may be found between scale degrees 1 to 4, 2 to 5, 3 to 6, 5 to 1, 6 to 2, and 7 to 3). Butler et al. also argue that the ordering of the notes of the tritone is important: a tritone F-B implies a tonal center of C much more strongly than B-F. Similarly, Vos (1999) has argued that a rising fifth or descending fourth at the beginning of a melody can be an important cue to key. These arguments are examples of what we earlier called a "structural" view of key perception. In support of such a view, some experiments have shown that the ordering of pitches does indeed have an effect on key judgments. Brown (1988) found, for example, that the pitches D-F#-A-G-E-C# elicited a strong preference for D major, whereas the sequence C#-D-E-G-A-F# was more ambiguous and yielded a judgment of G major slightly more often than D major (see also Auhagen, 1994; Bharucha, 1984; West & Fryer, 1990). Similarly, Matsunaga and Abe (2005) played participants tone sequences constructed from the pitch set {C, D, E, G, A, B} played in different orders. They found that the ordering affected key judgments, with certain orderings eliciting a strong preference for C major, some for G major, and some for A minor.[2]

While the studies of Brown (1988), Matsunaga and Abe (2005), and others might be taken to support the structural view of key perception, it would be a mistake to interpret them as refuting the distributional view altogether. For one thing, the sequences used in these studies are all extremely short; one might argue that such short sequences hardly provide listeners with enough "evidence" for a distributional strategy to be applied. Moreover, in some cases, the pitch sets used are deliberately constructed to be distributionally ambiguous. For example, the set {C, D, E, G, A, B} is fully contained in both the C major and G major scales, and also contains all three tonic triad notes of these two keys. The fact that structural cues are used by listeners in such ambiguous situations may have little relevance to real music, where distributional information generally provides more conclusive evidence as to key. We should note, also, that the "structural" view of key perception has yet to be worked out as a testable, predictive theory. It remains possible, however, that key perception depends significantly on the detection of certain structural musical patterns or on a combination of structural and distributional strategies.

As noted earlier, this question is difficult to resolve using real music, where both distributional and structural cues tend to be present. A better way to examine the role of distributional information would be to use melodies generated randomly from typical pitch-class distributions for different keys. In such melodies, the key would be indicated by the distribution, but it would presumably not be indicated by structural cues that depend on a particular temporal arrangement of pitches, such as a particular ordering of an interval, an implied harmonic progression, or the occurrence of certain scale degrees at particular points in the melody. If listeners are indeed relying on such structural cues, they may be unable to determine the underlying key and may even be misled into choosing another key.

Before continuing, we should briefly summarize other relevant studies that have explored listeners' sensitivity to pitch-class distribution. Several studies have employed a probe-tone methodology using musical materials quite different from those of Western tonal music. In a study by Castellano, Bharucha, and Krumhansl (1984), American participants were played passages of classical Indian music; probe-tone methods were used to see whether the responses reflected the distribution of pitch-classes in the input. Similarly, Oram and Cuddy (1995) and Creel and Newport (2002) did probe-tone studies using melodies generated from artificial pitch-class distributions designed to be very dissimilar to any major or minor scale. In all three of these studies, listeners' responses were highly correlated with the pitch-class distribution of the input—with tones occurring more frequently in the context being given higher ratings—suggesting that listeners are indeed sensitive to pitch-class distribution. We should not take these studies to

[2]One model that does not fit neatly into our structural/distributional taxonomy is Bharucha's (1987) neural-network model. This model consists of three levels of interconnected units representing pitches, chords, and keys; sounding pitches activate chord units which that in turn activate key units. The model is similar to distributional models in that it takes no account of the temporal ordering of pitches (except insofar as the activation of units decays gradually over time); however, the effect of pitches is mediated by the chords that contain them.

indicate that probe-tone responses in general are merely a reflection of the frequency of tones in the context (we return to this point below). But they do show that listeners are sensitive to pitch-class distribution, and this suggests that they might use distributional information in key identification as well.

A study by Smith and Schmuckler (2004) investigated the role of distributional information in key-finding. In this study, probability distributions were created using the Krumhansl-Kessler profiles, either in their original form or with the profile values raised to various exponents (in order to increase the degree of differentiation between tones in the profile). These distributions were used to control both the duration and the frequency of occurrence of pitches, which were then randomly ordered. Thus the experiment tested participants' ability to use distributional cues in the absence of structural ones. Participants were played these melodies, and their perceptions of key were measured using a probe-tone methodology. Profiles representing their responses were created, and these were correlated with Krumhansl and Kessler's probe-tone profiles. A high correlation with the K-K profile of a particular key was taken to indicate that participants heard the melody in that key. The authors found that listeners' judgments did indeed reflect perception of the correct key, especially when the key-profiles used to generate the melodies were raised to high exponents. The authors found that the total duration of each pitch-class in the melody is important; increasing the number of events of a certain pitch-class but making them shorter (so that the total duration of each pitch-class is the same) does not result in a clearer perception of tonality for the listener.

Smith and Schmuckler's (2004) study seems to point to a role for distributional information in key perception. However, it is open to two possible criticisms. The first concerns the fact that participants' judgments of key were measured by gathering probe-tone responses and correlating these with the original K-K profiles. This is a highly indirect method of accessing key judgments (see Vos, 2000, for discussion). It is true that probe-tone studies using a wide variety of tonal contexts have yielded quite consistent responses (Cuddy, 1997; Krumhansl, 1990); this suggests that probe-tone profiles are, indeed, a fairly reliable indicator of key judgments. But it is still possible that probe-tone responses are affected by the precise context that is used, at least to some extent. An alternative method, which has been used in some earlier studies of key perception (Brown, 1988; Cohen, 1991; Matsunaga & Abe, 2005), is to ask participants to report their key judgments directly. This "direct" method is impractical with untrained participants, who may be unable to articulate their knowledge of key, but with trained participants—as will be used in this study—this problem does not arise.

A second criticism concerns Smith and Schmuckler's (2004) analysis of their data. The authors indicate that, in some conditions at least, listeners' probe-tone responses to distributional melodies were highly correlated with the K-K profile for the correct key. But they do not indicate whether the K-K profile of the correct key was the *most* highly correlated with the probe-tone responses. If the profile of the correct key matched the probe-tone responses better than any other, this might be taken to indicate that the participants had judged the key correctly; but this information is not given. Thus, the results remain inconclusive as to whether listeners can judge key based on distributional information alone.[3]

In this study, we present an experiment similar to that done by Smith and Schmuckler (2004), but with three differences. First, the probability distributions used to create our melodies were generated from a musical corpus, rather than from experimental perception data (as in Smith and Schmuckler's study). Second, we measured participants' intuitions about key using explicit key judgments, rather than using the more indirect probe-tone method. Third, we measure the influence of pitch-class distribution on listeners' responses by looking at the proportion of key judgments that matched those predicted by the pitch-class distribution. In so doing, we compare several different distributional models of key-finding, to see which one achieves the best fit with the participants' responses. We consider the Krumhansl-Schmuckler model, Temperley's probabilistic model (described above), and several variants of the probabilistic model.

Finally, we examine the question of whether absolute pitch (AP) possession aids or hinders key-finding in distributional melodies. In general, the perception of key is assumed to be relative, not absolute. Most listeners

[3]We should note also that the distributions used to generate the melodies in Smith and Schmuckler's (2004) study were based on the K-K profiles. Since these profiles are drawn from perception data, one might question whether they really reflect the distribution of tones in tonal music. It is clear that the K-K profiles are qualitatively very similar to pitch-class distributions in tonal music—a comparison of Figures 2 and 3 demonstrates this. Quantitatively, they are not so similar (even when normalized to sum to 1), as the values for chromatic pitches are much too high; some kind of nonlinear scaling is needed to adjust for this, as seen in Smith and Schmuckler's study. An alternative approach would be to generate the melodies using distributions drawn from actual music, as we do in the current study.

cannot listen to a melody and say "that is in C major"; rather, they identify the key by recognizing that a particular note is the tonic pitch and that the melody is in major or minor. A small fraction of the population—those with absolute pitch—are able to identify pitches (and therefore keys) in absolute terms (for an overview, see Levitin & Rogers, 2005; Takeuchi & Hulse, 1993; Terhardt & Seewann, 1983). Based on earlier research on absolute pitch (Marvin, 1997), we hypothesized that participants with absolute pitch might differ in their key-finding strategy from those with relative pitch—perhaps identifying key in a more deliberate and methodical way, even explicitly counting pitches to determine a distribution. To test this, we grouped participants according to their absolute pitch ability, and tested the groups in both "timed" and "untimed" conditions. In Experiment 1 (the untimed condition), participants heard the entire melody and then made a key judgment; in Experiment 2 (the timed condition), they stopped the melody when they felt they had identified the key, and then reported their judgment. The stimuli in Experiments 1 and 2 were different, but were generated by the same algorithm. Our hypothesis was that listeners with absolute pitch might use a more deliberate "counting" strategy to determine the key, and therefore might take more time to reach a judgment than those with relative pitch.

## Method

### *Participants*

Data are reported here for 30 participants (18 male, 12 female) with a mean age of 19.08 years ($SD = 0.97$), who volunteered to take part in both experiments and were paid $10 for participating. All were undergraduate music students at the Eastman School of Music of the University of Rochester. Participants began studying a musical instrument at a mean age of 7.65 years ($SD = 3.57$), and thus had played for more than 11 years. All participants had completed at least one year of collegiate music theory study. Twenty-one participants identified themselves as Caucasian, seven as Asian, one as Hispanic, and one as African-American.

Although we initially asked participants to report their status as AP or non-AP listeners, we administered an AP posttest to all participants to confirm. Interestingly, the distribution of scores was trimodal, with high- and low-scoring groups and a distinct group of scores in the middle. Based on the distribution of scores, those who scored 85% or higher ($M = 97\%$, $n = 12$) we classified as AP; those who scored 25% or lower ($M = 10\%$, $n = 11$) we classified as non-AP; and participants with scores between 40% and 60% ($M = 53\%$, $n = 7$), we classified as "quasi-AP." Of the seven quasi-AP participants, two self-identified as AP, two as non-AP, and three as quasi-AP.[4] AP participants began their instrumental training at age 6.2 years, non-AP at 8.8 years, and quasi-AP at 8.1 years. All seven Asian participants placed either in the AP or quasi-AP group, and the first language of five of the seven was Mandarin, Cantonese, or Korean (see Deutsch, Henthorn, Marvin, & Xu, 2006; Gregersen, Kowalsky, Kohn, & Marvin, 2001). Of the AP and quasi-AP participants, all but two played a keyboard or string instrument. Of the non-AP participants, none played a keyboard instrument, two played a string instrument, and one was a singer; the majority ($n = 7$) played woodwind and brass instruments.

### *Apparatus*

Two experiments were administered individually to participants in an isolated lab using a custom-designed program in Java on an iMac computer, which collected all responses and timings for analysis. All participant responses were made by clicking on-screen note-name buttons with the mouse. Stimuli were presented via BeyerDynamic DT770 headphones, and participants had an opportunity to check note names on a Kurzweil PC88mx keyboard next to the computer before completing each trial. Before beginning the experiment, participants were given an opportunity to adjust the loudness of sample stimuli to a comfortable listening level.

### *Stimuli*

Simuli for both experiments consisted of melodies generated quasi-randomly from scale-degree distributions. The distributions were created from a corpus consisting of the first eight measures of each of the string quartet movements by Mozart and Haydn.[5] The pitches of each

[4]Responses for three of the "quasi-AP" participants, when asked whether they had AP, were "sort of" and "I don't think so, but my teaching assistant does." One quasi-AP bassoon player wrote that he has AP only for the "bottom half of the piano."

[5]The corpus was taken from the Musedata archive (www.musedata.org). The archive contains the complete string quartets of Mozart (78 movements) and Haydn (232 movements) encoded in so-called "Kern" format (Huron, 1999), representing pitches, rhythms, bar lines, key symbols (indicating the main key of each movement), and other information. It was assumed that very few of the movements would modulate before the end of the first eight measures; thus, in these passages, the main key of the movement should also generally be the "local" key.

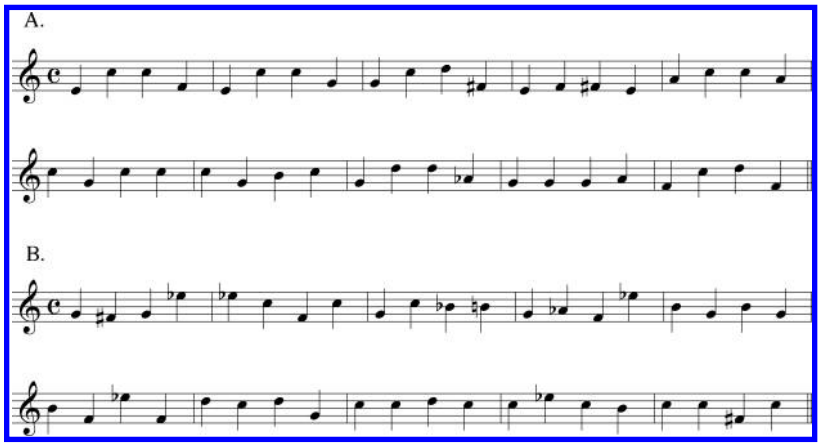


FIGURE 4. Two melodies used the experiments. Melody A, with a generating key of C major, was used in Experiment 1; Melody B, with a generating key of C minor, was used in Experiment 2.

8-measure passage were converted into scale degrees in relation to the main key of the movement. A scale-degree profile, showing the proportion of events of each scale degree, was then created for each passage. (These profiles reflected only the *number* of events of each scale degree, not their duration.) The profiles of all major-key passages were averaged to create the major key-profile (giving each passage equal weight), and the same was done for minor-key passages. This led to the profiles shown in Figure 3. It can be seen that the profiles in Figure 3 are qualitatively very similar to the Krumhansl-Kessler profiles shown in Figure 2 (recall that the K-K profiles were generated from experimental probe-tone data). Both profile sets reflect the same three-level hierarchy of tonic-triad notes, scalar notes, and chromatic notes. (One difference is that in the minor-key Mozart-Haydn profile, 7 has a higher value than b7, while in the Krumhansl-Kessler profiles the reverse is true; thus the Mozart-Haydn profiles reflect the "harmonic minor" scale while the K-K profiles reflect the "natural minor.")

The profiles in Figure 3 were used to generate scale degrees in a stochastic fashion (so that the probability of a scale degree being generated at a given point was equal to its value in the key-profile). Each melody was also assigned a randomly chosen range of 12 semitones (within an overall range of A3 to G5), so that there was only one possible pitch for each scale degree. Using this procedure, we generated 66 melodies (30 for each experiment, and six additional for practice trials), using all 24 major and minor keys, each one 40 notes in length. Figure 4 shows two of the melodies, generated from the key-profiles for C Major and C minor. The melodies were isochronous, with each note having a duration of 250 ms, and were played using the QuickTime 7.1.2 piano timbre.

Stimuli for the AP posttest were those of Deutsch, Henthorn, Marvin, and Xu (2006), used with permission. Participants heard 36 notes spanning a three-octave range from $C_3$ (131 Hz) to $B_5$ (988 Hz). The notes were piano tones generated on a Kurzweil synthesizer and played via computer MP3 file. To minimize the use of relative pitch as a cue, all intervals between successively presented notes were larger than an octave.[6]

### *Procedure*

Participants took part in two experiments in a single session, with a rest between. Before each experiment, participants heard three practice trials and were given an opportunity to ask questions and adjust the volume; no feedback was given. In Experiment 1, participants heard 30 melodies as described above; the computer program generated a new random order for each participant. Pacing between trials was determined by the participant, who clicked on a "Play" button to begin each trial. After hearing each stimulus melody, the participant was permitted (but not required) to sing or whistle his/her inferred tonic and then to locate this pitch on the keyboard in order to determine the pitch name. (This step was largely unnecessary for AP participants, but they were given the same opportunity to check their pitch names at the keyboard.) Participants then clicked on one of 12 buttons (C, C#/Db, D, D#/Eb, E, and so on) to register their tonic identification. A second screen asked them to click on "major" or "minor" to register the perceived mode of the melody. Experiment 2

[6]In scoring the AP posttest, we permitted no semitone deviations from the correct pitch label, as is sometimes done in scoring such tests.

was identical in format, except that the participants heard 30 new melodies (generated in the same manner), and were urged to determine the tonic and mode as quickly as possible. When the participant could sing or hum a tonic, he/she clicked on a button that stopped the stimulus and a response time was collected at that point. Then extra time could be taken with the keyboard to determine the note name and enter the participant's response.

After the two experiments, participants took an AP posttest. Pitches were presented in three blocks of twelve, with 4-s intervals between onsets of notes within a block, and 30-s rest periods between blocks. Participants were asked to write the letter name of each pitch on a scoring sheet (no octave designation was required). The posttest was preceded by a practice block of four notes. No feedback was provided, either during the practice block, or during the test itself. Students were not permitted to touch the keyboard for the posttest. Finally participants filled out a questionnaire regarding their age, gender, training, and AP status, as well as the strategies they employed in completing the experimental tasks.

## Results

The main question of interest in our experiments is the degree to which participants' key judgments accorded with the keys used to generate the melodies—what we will call the "generating" keys.[7] Before examining this, we should consider whether it is even *possible* to determine the generating keys of the melodies. This was attempted using Temperley's probabilistic key-finding model, described earlier. Using the key-profiles taken from the Haydn-Mozart corpus (the same profiles used to generate the melodies), this model chose the generating key in all 60 melodies used in the experiment. This shows that it is at least computationally possible to identify the generating key in all the melodies of our experiment using a distributional method.

Turning to the participant data, our 30 listeners each judged the key of 60 melodies: 30 in Experiment 1 (untimed) and 30 in Experiment 2 (timed). This yielded 900 data points for each of the two experiments and 1800 data points in all. Comparing participants' judgments to the generating keys, we found that .51 ($SE = .03$) of the judgments matched the generating key in the untimed experiment and .52 ($SE = .03$) in the timed experiment. For each participant, the mean proportion correct was calculated, and these scores were compared with a chance performance of 1/24 or 4.2% (since there are 24 possible keys), using a one-sample *t*-test (two-tailed). We found performance to be much better than chance on both the untimed experiment, $t(29) = 17.71$, $p < .0001$, and the timed experiment, $t(29) = 14.89$, $p < .0001$.

We then examined the amount of agreement between participants. For each melody, we found the key that was chosen by the largest number of participants—we will call this the "most popular key" (MPK) for the melody. The MPK judgments matched the generating keys in 50 out of the 60 melodies.[8] Overall, the MPK judgments accounted for only 56.1% of the 1800 judgments. This is an important result for two reasons. First, it is surprisingly low: One might expect general agreement in key judgments among our participants, who are highly trained musicians. But with these melodies, the most popular key choices only accounted for slightly more than half of the judgments. We return to this point later. Second, as we try to model listeners' judgments in various ways, we should bear in mind that no model will be able to match more than 56.1% of the judgments in the data. (One cannot expect a model to match 100% of participants' judgments when the participants do not even agree with each other.)

As a second way of measuring agreement among listeners, we calculated the Coefficient of Concentration of Selection (CCS) for the responses to each melody (Matsunaga & Abe, 2005). The CCS is a measure of the level of agreement on a categorical response task, and is defined as

$$\mathrm{CCS} = (\chi^2/\{N(K-1)\})^{1/2} \tag{3}$$

where $\chi^2$ is the chi-square of the distribution of responses, $N$ is the number of responses, and $K$ is the number of response categories. The CCS varies between 0 (if responses are evenly distributed between all categories) and 1 (if all responses are in the same category). For our melodies, the CCS values ranged from .31 to 1.00;

[7] We do not call them the "correct" keys, because the correct key of a randomly generated melody is a problematic notion. Suppose the generative model, using the key of C major, happened to generate "Twinkle Twinkle Little Star" in F# major (F# F# C# C# . . . )—which could happen (albeit with very low probability). Would this mean that the correct key of this melody was C major? Surely not. It seems that "correct" key of such a melody could only be defined as the one chosen by listeners.

[8] To be more precise: In 48 of the 60 cases, there was a single most popular key that was the generating key. In two other cases, two keys were tied for most popular, but in both of these cases one of the two keys was the generating key. For simplicity, we counted the generating key as the most popular key in those two cases.

the average across our 60 melodies was .589.[9] As a comparison, Matsunaga and Abe (2005) provide CCS values for the 60 six-note melodies used in their experiment; the average CCS value for these melodies was .52.

We then considered the question of whether possessors of absolute pitch performed differently from other listeners. With regard to matching the generating keys, on the untimed experiment the AP participants achieved an average score of .56 correct ($SE = .051$), the quasi-AP participants achieved .42 ($SE = .03$), and the non-AP participants achieved .51 ($SE = .04$); the difference between the groups was not significant, $F(2, 27) = 2.29$, $p > .05$. On the timed experiment, too, the mean scores for the AP participants ($M = .54$, $SE = .05$), the quasi-AP participants ($M = .51$, $SE = .07$), and the non-AP participants ($M = .49$, $SE = .05$) did not significantly differ, $F(2, 27) = 0.21$, $p > .05$. We then examined the average time taken to respond on the timed experiment; we had hypothesized that AP participants might use an explicit counting strategy and therefore might take longer in forming key judgments. The AP participants showed an average time of 7.09 ($SE = 0.34$) seconds, the quasi-AP participants yielded an average time of 7.45 ($SE = 0.43$) seconds, and the non-AP participants yielded an average time of 7.16 ($SE = 0.55$) seconds. (On average, then, the AP and non-AP participants heard about 27 notes of each melody and the quasi-AP participants heard about 28 notes.) The difference between the three groups was not significant, $F(2, 27) = 0.14$, $p > .05$. Thus, we did not find any significant difference between AP, quasi-AP, and non-AP participants with regard to either the speed of their judgments or the rate at which they matched the generating keys.

## Discussion

The experiments presented above were designed to examine whether listeners are able to identify the key of a melody using distributional information alone. The results suggest that listeners can, indeed, perform this task at levels much greater than chance. This result was found both in an untimed condition, where the complete melody was heard, and in a timed condition, where participants responded as quickly as possible. However, only slightly more than half of participants' judgments matched the generating key, both in the timed and the untimed conditions. No significant difference in key-finding performance was found with regard to absolute pitch.

One of our goals in this study was to test various distributional models of key-finding to assess how well they matched listener judgments. In what follows, we begin by examining the performance of the Temperley probabilistic model described earlier; we then consider several other models and variants of this model. One issue to consider here is the distinction between timed and untimed conditions. In the untimed condition, listeners heard the entire melody before judging the key; in the timed condition, they generally did not hear the entire melody. (As noted above, participants on average heard about 27 notes, or about two thirds, of the melody in the timed condition. In only 199 of the timed trials, or 22.1% of the total, did participants "run out the clock" and hear the entire melody.) It seems questionable to compare the judgment of a model that had access to the entire melody with that of a listener who only heard part of the melody; on the other hand, participants did hear *most* of the timed melodies, and adding in the timed melodies provides a larger body of data. For the most part, we focus here on the untimed melodies, but in some cases we consider both untimed and timed melodies; this will be explained further below.

One simple way of testing a key-finding model against our data is by comparing its key judgments to the MPK judgments—the keys chosen by the largest number of participants. We noted above that the MPK judgments matched the generating key on 50 out of 60 melodies (considering both untimed and timed melodies), and the Temperley probabilistic model matched the generating key on all 60 melodies. Thus the Temperley probabilistic model matches the MPK judgments in 50 out of 60 melodies. On the untimed melodies, the Temperley model matched 26 out of 30 MPK judgments. (See the first row of Table 1.)

We also considered two other measurements of how well the model's output matched the participants' judgments. One measure makes use of the fact that the probabilistic model calculates a probability for each key given the melody, the key with the highest probability being the preferred key. The model's probability for the generating key, what we will call $P(K_g)$, can be used as a measure of the model's "degree of preference" for that key. The participants' degree of preference for the generating key can be measured by the number of responses that key received, or *responses*$(K_g)$. If the probabilistic model is capturing participants' key judgments, then the probability it assigns to the generating key should be

[9] We also wondered if the CCS was lower on melodies for which the MPK was not the generating key. For the 10 melodies on which the MPK was not the generating key, the average CCS was .48; for the other 50 melodies, the average CCS was .61.

**TABLE 1. Comparison of Key-Finding Algorithms.**

| Model | No. of matches to generating keys (60 melodies) | No. of matches to MPKs (untimed condition only, 30 melodies) | No. of matches to MPKs (untimed and timed conditions, 60 melodies) | Spearman correlation coefficient between rankings of keys by participants and model (averaged over 30 untimed melodies) |
|---|---|---|---|---|
| Probabilistic model (PM) | 60 (100.0%) | 26 (86.7%) | 50 (83.3%) | .54 |
| PM with Essen profiles | 58 (96.6%) | 25 (83.3%) | 48 (80.0%) | .53 |
| Krumhansl-Schmuckler model | 49 (81.7%) | 23 (76.7%) | 43 (71.7%) | .45 |
| PM ignoring last 20 notes | 52 (86.7%) | 22 (73.3%) | 45 (75.0%) | .52 |
| PM ignoring first 5 notes | 59 (98.3%) | 27 (90.0%) | 51 (85.0%) | .53 |
| PM favoring major-mode keys ($mf = .999$) | 59 (98.3%) | 25 (83.3%) | 49 (81.7%) | .53 |
| "First-order" probabilistic model | 56 (93.3%) | 27 (90.0%) | 49 (81.7%) | .49 |
| PM with profile value for tonic multiplied by 1000 on first note | 59 (98.3%) | 26 (86.7%) | 51 (85.0%) | .55 |

higher in cases where more participants chose that key. One problem is that, for the 30 untimed melodies, $P(K_g)$ varies beween 0.98 and 0.9999999; the variation in these numbers is not well captured either by a linear scale or a logarithmic scale. A better expression for this purpose is $\log(1- P(K_g))$; if this value is low, that means the model strongly preferred the generating key. (For our melodies, this varies between a low of −17.55 and a high of −4.01.) These values were calculated for each of the untimed melodies; Figure 5 plots $\log(1 - P(K_g)$ against $responses(K_g)$ for each melody. The observed relationship is in the predicted direction (in cases where $\log(1 - P(K_g))$ is lower, the generating key received more responses); however, it is very small and statistically insignificant ($r = .24$). Thus, by this measure, the Temperley model does not fare very well in predicting participants' degree of preference for the generating key.

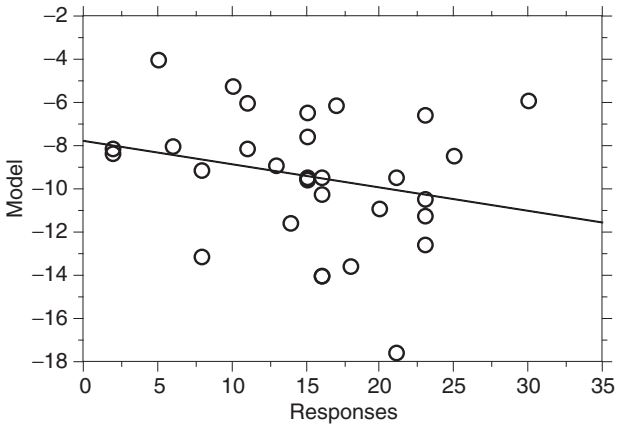


FIGURE 5. The model's degree of preference for the generating key of each melody ($\log(1 - P(K_g))$ (vertical axis) plotted against the number of responses for that key (horizontal axis), for the 30 melodies in the untimed experiment.

The two measures used so far both consider only the most preferred keys of the model and the participants. It would be desirable to compare the degree of preference for lower-ranked keys as well. Here we use Spearman's Rank Correlation, which correlates two rankings for a set of items without considering the numerical values on which those rankings were based. For each of the untimed melodies, we used the log probabilities generated by the Temperley probabilistic model for each key, $\log(P(\text{key} \mid \text{melody})$, to create a ranking of the 24 keys; we then used the participant data to create another ranking, reflecting the number of votes each key received (keys receiving the same number of votes were given equal rank). For each melody, we calculated the Spearman coefficient between these two rankings; for the 30 untimed melodies, the average correlation was .539.

Figure 6 shows two of the melodies in our experiment, and Figure 7 shows data for them: the number of votes for each key and the model's probability judgment for each key, $\log(P(\text{key} \mid \text{melody})$. For the first melody (with a generating key of D major), the participant responses are fairly typical, both in the degree of participant agreement (CCS = .50) and the number of votes for the generating key (15). The second melody (with a generating key of Eb minor) is the one that yielded the

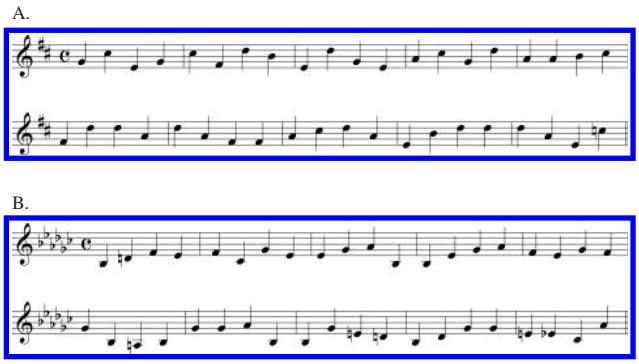


FIGURE 6. Two melodies used in the untimed experiment.

*minimum* number of votes for the generating key; this key received only two votes. Comparing the model's judgments to the participant judgments for these two melodies, we see that the fit is far from perfect; still, there is clearly some correspondence, in that the peaks in the participant data (the keys receiving votes) generally correspond to peaks in the model's values, especially in the first melody. Perhaps the most striking difference is that on the second melody, Bb major received the highest number of participant votes (8) but received a fairly low score from the model. The reason for the model's low score for Bb major is clear: there are 16 notes in the melody that go outside the Bb major scale. As to why the participants favored Bb major, perhaps the Bb major triad at the beginning of the melody was a factor (though bear in mind that only 8 out of 30 participants voted for this key). We will return below to the issue of what other factors besides pitch-class distribution may have affected the participants' judgments.

We now consider whether any other model can be found that achieves a better "fit" to the participant data than the Temperley probabilistic model. Table 1 shows the results for various models. First we show the number of matches between the model's judgments and the generating keys (for untimed and timed melodies combined). We then show the number of matches between the model's preferred keys and the MPK judgments. (We give results for the 30 untimed melodies; since this offers only a small body of data, we also give results for the timed and untimed melodies combined.) Finally, we show the Spearman correlation calculated between the rankings of keys by the participants and the model (for the untimed melodies only).

In considering alternatives to the Temperley probabilistic model, we first wondered how much the model's judgments were affected by the specific key-profiles that it used. To explore this, the model was run with a set of profiles gathered from another corpus. The corpus used was the Essen folksong database, a corpus of 6,200 European folk songs, annotated with pitch and rhythmic information as well as key symbols.[10] The Essen profiles (shown in Figure 8) are very similar to the Mozart-Haydn profiles (Figure 3), though with a few subtle differences. (In the Essen profiles, b7 has a higher value than 7 in minor, like the Krumhansl-Kessler profiles and unlike the Mozart-Haydn profiles.) Using the Essen profiles, the model matched the generating keys in 58 out of 60 cases (as opposed to 60 out of 60 with the Mozart-Haydn profiles). Thus the model's identification of generating keys does not seem to depend heavily on the precise values of the profiles that are used. The model's judgments when using the Essen profiles matched the MPK judgments in 48 out of 60 cases. This suggests that, in modelling listeners' distributional knowledge of tonal music, classical string quartets and European folk songs are almost equally good, though classical string quartets may be marginally better.

The next model tested was the Krumhansl-Schmuckler model. As discussed earlier, the K-S model operates by creating an "input vector" for the piece—a profile showing the total duration of all 12 pitch-classes in the piece; the correlation is calculated between this vector and the 24 K-K key-profiles, and the key is chosen yielding the highest correlation. Unlike the profiles of the probabilistic model, which were set from a corpus of music, the profiles of the K-S model were gathered from experimental data on human listeners (Krumhansl, 1990).[11] Thus, one might expect the K-S model to match our participants' judgments better than the probabilistic model. In fact, however, the K-S model yielded a poorer match to our listener data, matching only 43 of the 60 MPK judgments. The K-S model also fared worse at matching the generating keys; it matched only 49 out of 60 generating keys, whereas the probabilistic model matched all 60.

One suprising aspect of our experimental data is that participants matched the generating key at almost the same rate in the timed condition (where they made a key judgment as soon as they were able) as in the untimed condition (where they heard the entire melody).

[10]The Essen database is available at <http://kern.ccarh.org/cgi-bin/ksbrowse?l=/essen/>. It was created by Schaffrath (1995) and computationally encoded in "Kern" format by Huron (1999).

[11]In fact, the participants in Krumhansl and Kessler's (1982) study were rather similar to those of our experiment, namely undergraduates with high levels of music training. However, while Krumhansl and Kessler's subjects generally did not have training in music *theory*, most of our participants had studied collegiate music theory for three semesters and therefore did have some theory background.

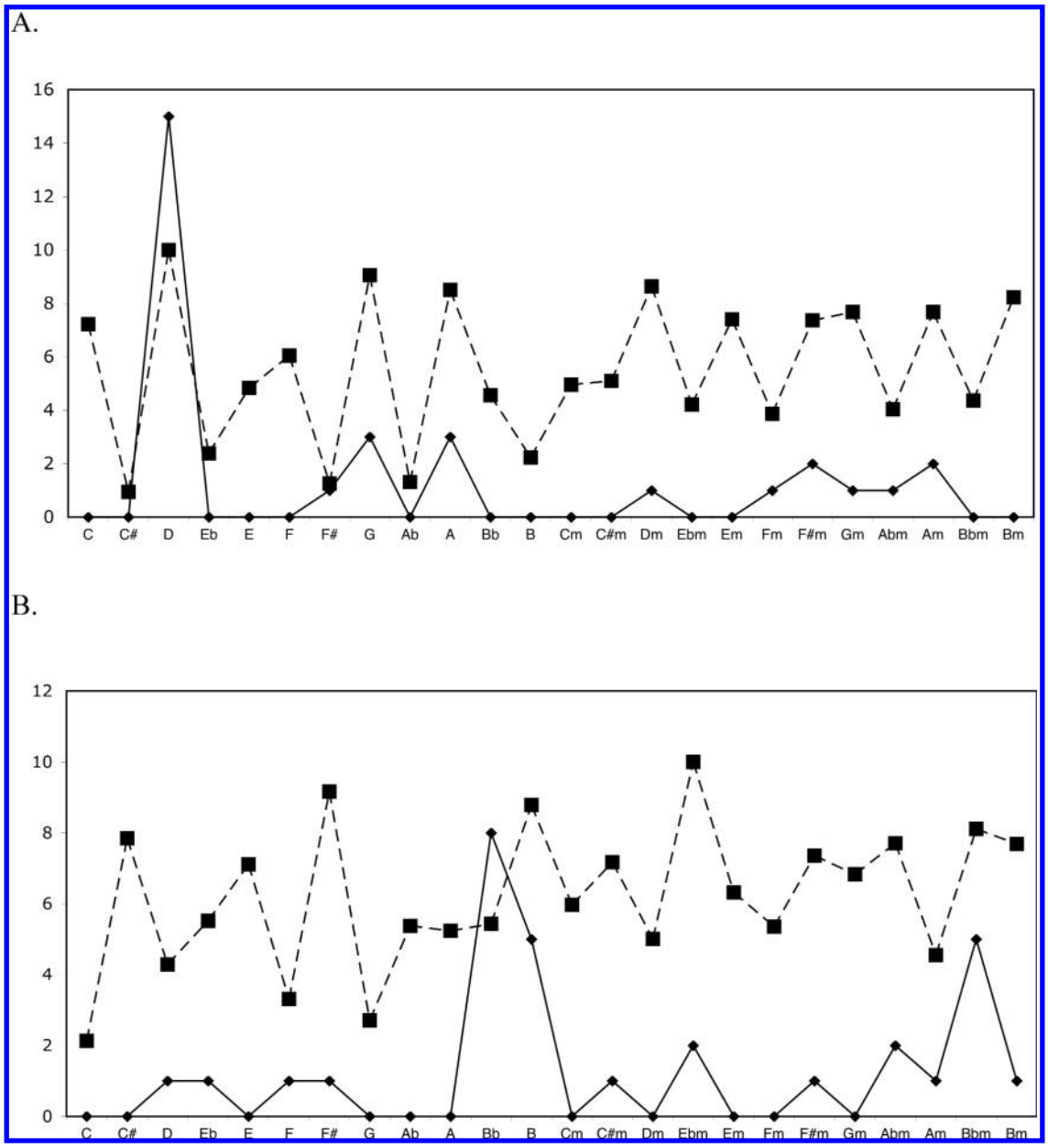


FIGURE 7. Data for the melodies shown in Figure 6A (above) and Figure 6B (below). The solid line shows the number of votes for each key; the dotted line shows the probabilistic model's score for that key, log(*P*(key | melody). (The model's scores have been normalized to allow comparison with participant responses.)

This suggested to us that perhaps participants were using a distributional strategy, but basing their judgment only on the first portion of the melody. In our own experience of the melodies, too, we felt that we sometimes gave greater weight to notes early in the melody—perhaps forming a key judgment after just a few notes and then fitting any subsequent notes into that key framework. Thus, we reasoned that the fit of the probabilistic model to the participants' judgments might be improved by simply running it on only the first portion of the melody—ignoring the last *n* notes. This approach was tried, for various values of *n*; however, no improvement in fit to the listeners' responses could be achieved in this way. Table 1 shows the results for $n = 20$ (a number of other values were also tried). Using only the first 20 notes of each melody, the model

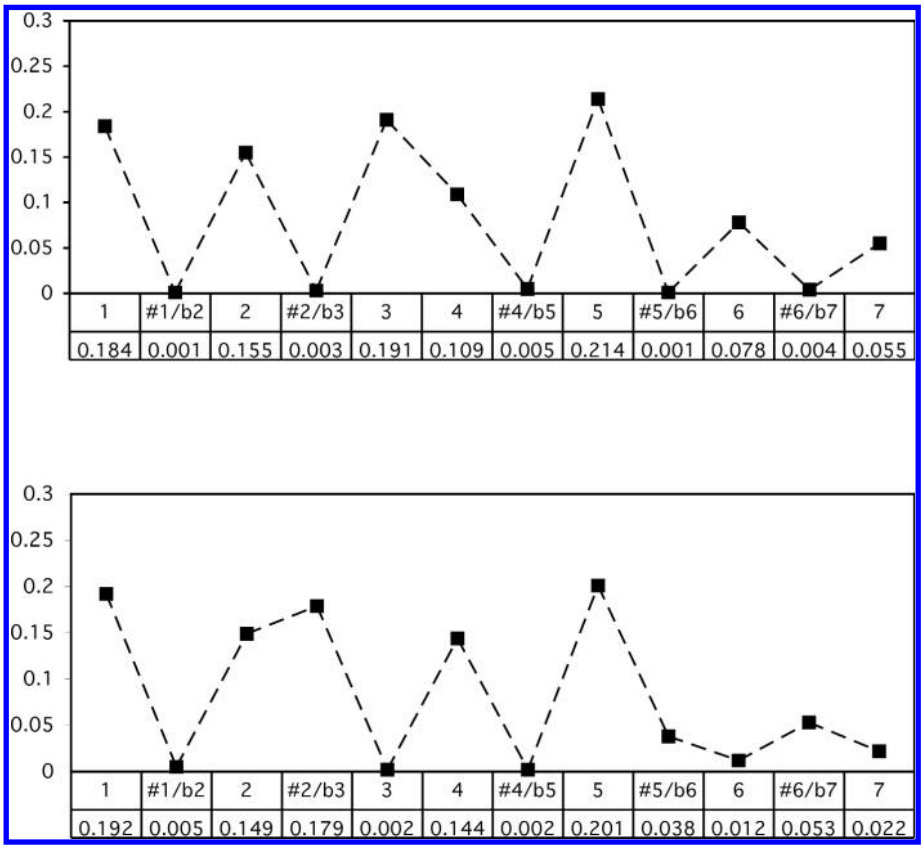


FIGURE 8. Key-profiles generated from the Essen Folksong Collection, for major keys (above) and minor keys (below).

matched only 45 out of the 60 untimed and timed MPKs, as opposed to 50 when the entire melody was considered. We then tried ignoring notes at the beginning of the melody—reasoning that perhaps participants gave greater weight to notes near the end. (An effect of final position was observed by Creel & Newport, 2002.) The model was modified to ignore the first *n* notes, using various values of *n*. Again, little improvement could be obtained. With $n = 5$ (ignoring the first 5 notes of the melody), the number of matches to the MPKs on the combined untimed and timed melodies increased from 50 to 51; other values of *n* resulted in fewer matches. In short, there is little evidence that participants considered only part of the melodies in making their judgments; in general, the model seems to match listeners' judgments most closely when it considers the entire melody.

Another issue concerns possible biases in the key judgments. The probabilistic model used here assumes that all keys are equal in prior probability; but this may not be the assumption of listeners. Certainly some tonics are more common than others—there are more pieces in C major than in F# major, and generally, it seems clear that "white-note" keys are more common than "black-note" keys. Perhaps these prior probabilities are reflected in listeners' judgments. This possibility is of particular interest with regard to AP participants. Other studies have found a "white-note" bias in AP listeners with regard to pitch identification—such listeners can identify white-note pitches more rapidly than black-note pitches (Marvin & Brinkman, 2000; Miyazaki, 1989, 1990; Takeuchi & Hulse, 1991); we wondered if a similar bias would affect their key judgments. (Although one might not expect such a bias with non-AP listeners, Marvin and Brinkman found that non-AP listeners also reflected a white-note bias in pitch identification.) To address this question, we examined the frequency of white-note versus black-note key judgments, considering just the data from the untimed experiment.

Out of the 30 melodies in the untimed experiment, 12 had generating keys with "black-note" tonics; thus, if there was no bias, we would expect the proportion

of responses for black-note tonics to be 12/30 = .4. For the AP group, the actual proportion of responses for black-note tonics (averaged across participants) was exactly .4 (*SE* = .015), which obviously did not significantly differ from the expected value, $t(11) = 0.00$, $p > .05$. The proportion of responses for black-note tonics for the quasi-AP and non-AP groups also did not differ significantly from the expected value. Thus we find no evidence for any bias towards white-note tonics.

We then considered whether listeners might have a bias regarding mode. Again, it seems likely that major mode is more common than minor mode in most tonal styles, and this might affect listeners' judgments. Out of the 30 melodies in the untimed experiment, 15 had minor-mode generating keys, thus we would expect the proportion of votes for minor-mode keys to be .50. For both the non-AP group and the AP group, the observed proportion of votes for minor-mode keys was significantly less than the expected value. For the non-AP group, the mean was .397, *SE* = .035, $t(10) = -2.95$, $p < .05$; for the AP group, the mean was .401, *SE* = .020, $t(11) = -4.62$, $p < .001$. For the quasi-AP group, the observed proportion was .529, not significantly different from the expected value, $t(6) = 0.56$, $p > .05$. Combining the three groups together, we find a mean proportion of .431 (*SE* = .021) for minor-mode keys, reflecting a significant bias towards major-mode keys, $t(29) = -3.27$, $p < .01$.

We then tried incorporating this bias into the model. Recall that our original expression for $P(\text{key} \mid \text{melody})$ (equation 1) contained a factor representing the prior probability of the key—$P(\text{key})$—but that this was assumed to be equal for all keys and therefore neglected. We now assume that $P(\text{key})$ may be different for different keys. The probability of each major key can be expressed as *mf*/12, where *mf* (*mode factor*) is the probability of a major-mode form of a key. The probability of the parallel minor is then (1–*mf*)/12; this ensures that the probabilities for all 24 keys sum to 1. A high value of *mf* (close to 1) would give major keys a higher prior probability than minor keys. Various values of *mf* were tried, but none yielded any improvement over the original model in terms of matching the MPK judgments. With *mf* = .9, the model matched 50 of the untimed and timed MPKs (the same as the original model); with *mf* = .999, it matched 49 of the MPKs.

We then considered a distributional model of a rather different kind. Experiments have shown that listeners are sensitive to transitional probabilities between adjacent surface elements, both in language and in music (Saffran & Griepentrog, 2001; Saffran, Johnson, Aslin, & Newport, 1999). We wondered if this was the case with regard to key perception as well.[12] Specifically, we considered a generative model again based on the Mozart-Haydn corpus, in which the probability of a note $N_n$ at a particular point depends on the key and on the previous note, $N_{n-1}$; the probability of an entire melody is then the product of these probabilities over all notes:

$$P(\text{melody} \mid \text{key}) = \Pi_n P(N_n \mid N_{n-1}, \text{key}) \qquad (4)$$

(One could call this a "first-order Markov model"—conditional on the key—whereas our original model was a "zeroth-order Markov model.")[13] Again, using Bayesian logic, this expression is proportional to the probability of a key given the melody. The probabilities were set using the Mozart-Haydn corpus, by finding the probabilities of all "scale-degree transitions"—the probability of each scale degree, given the previous scale degree (within the same line). In a sense, this could be considered a kind of "structural" model, as it considers the ordering of notes (in a very limited way). In particular, it seemed that this model might improve the model's sensitivity to chromatic notes—notes outside the scale, which almost invariably resolve by a half-step. Incorporating scale-degree transitions gives the model the ability to capture this convention; it should know, for example, that a note that is not followed by half-step is unlikely to be chromatic and almost certainly a note of the scale.

As Table 1 shows, this "first-order" model once again failed to yield any improvement over the original probabilistic model. Considering both the timed and untimed melodies, it matched the MPKs on 49 out of the 60 melodies (though intriguingly, it got one point more than the original model on the untimed melodies). The model also performed worse than the original model with regard to the generating keys, matching only 56 out of 60. Thus, we find that a probabilistic model based on note transitions matches human judgements of key no better than the original model.

---

[12]Earlier attempts to incorporate such information into key-finding models have had limited success. In particular, Toiviainen and Krumhansl (2003) incorporated information about note transitions—the probability of one note following another—into a distributional model, but found that it yielded no improvement over the Krumhansl-Schmuckler model. In that case, however, the transitional information used was gathered from perception data regarding the perceived "relatedness" of tones; in the present case, we set the transitional probabilities based on a musical corpus.

[13]For the first note, the probability cannot be calculated in this way, since there is no previous note; in this case, then, the "zeroth-order" profiles were used.

**TABLE 2. Ten Melodies with Participant Responses.**

| Melody (our code name) | Generating key | Most popular key | Second most popular key | CCS |
|---|---|---|---|---|
| mcsm1 | C# | Ab | C# | .55 |
| mcsn1 | C#m | Ab | C#m | .42 |
| mebn1 | Ebm | Bb | Bbm | .33 |
| mebn2 | Ebm | F# | B | .46 |
| me-m3 | E | B | E | .73 |
| mabm3 | Ab | C# | Ab | .77 |
| mbbm3 | Bb | Gm | Eb | .43 |
| mcsn3 | C#m | Abm | F#m/C#m (tie) | .31 |
| mabn3 | Abm | Ab | Abm | .32 |
| mg-n4 | Gm | Bb | Gm | .49 |

To gain further insight into participants' key judgments, we inspected the 10 melodies on which the MPK judgment disagreed with the generating key, to see if any other patterns emerged that might explain these judgments. We hypothesized that the MPK tonic might tend to be the last note of the melody; this was the case in only 4 of the 10 melodies. We also considered whether the MPK tonic was the first note of the melody; this was the case in 8 of the 10 melodies. This suggested that perhaps participants were using a "first-note-as-tonic" cue. However, inspection of the other melodies showed that the first note was the tonic of the most popular key on only 22 of the 60 melodies.[14] (We also tried incorporating this factor into the model, by giving it a special "first-note key-profile" in which the tonic value was especially high. This produced only a very slight improvement; as shown in Table 1, multiplying the tonic profile value by 1000 yielded a gain of one point in matching the untimed and timed MPKs.) We also examined the number of chromatic notes with respect to the MPK, thinking that perhaps listeners simply chose the key that yielded the minimum number of chromatic notes. But in 8 out of the 10 melodies on which the MPK judgment differed from the generating key, the number of chromatic notes in relation to the MPK was actually *higher* than the number in relation to the generating key. Thus we could not find any convincing explanation for why the MPK judgments might have differed from the generating key on these 10 melodies.

Out of all of the models we tried, none achieved more than a marginal improvement over the basic probabilistic model. Further insight into this is provided by the Spearman correlation measures, shown in Table 1; recall that these indicate the rank correlation between the model's "scores" for each key and the number of votes for that key in the participant data, averaged over the 30 untimed melodies. (For the probabilistic models, the score for each key is $\log(P(\text{key} \mid \text{melody})$; for the K-S model, it is the correlation value between the K-K key-profile and the input vector.) The basic probabilistic model yields a correlation of .539; the only model that achieves any improvement at all over this is the "first-note-as-tonic" model, which yields a correlation of .548.

A final issue concerns the keys that participants chose when they did not choose the generating key. One might expect, in such cases, that the key chosen would generally be closely related to the generating key. If we examine the 10 cases where the MPK differed from the generating key, we find that this is indeed the case (see Table 2). In the 4 cases where the generating key was major, the MPK was always either the major dominant (2 cases), the major subdominant (1 case), or the relative minor (1 case); in the 6 cases where the generating key was minor, the MPK was either the major dominant (2 cases), relative major (2 cases), parallel major (1 case), or minor dominant (1 case). The sample melodies shown in Figures 6 and 7 offer some insight into this as well; for both melodies, the most favored

[14]Some kind of "first-note" or primacy factor might also be suspected in the second melody in Figure 6; that melody begins by outlining a Bb major triad, which was the subjects' most favored key (see Figure 7). But the tests reported here suggest that neither a model which gives special weight to the first few notes of a melody, nor a model which strongly favors the first note as tonic, yields an improved fit to subjects' key judgments. Despite these findings, the possibility of some kind of first-note strategy cannot be ruled out. It may be, for example, that listeners use a distributional strategy initially, but when this yields an uncertain judgment they rely on a first-note strategy.

keys are closely related to the generating key.[15] A complete analysis of participants' key choices in relation to the generating key would be of interest, but this would lead us into the complex topic of key relations (Krumhansl, 1990; Lerdahl, 2001) and is beyond the scope of this study.

## Conclusions

Three general conclusions emerge from our study. First, when asked to determine the key of melodies generated from pitch-class distributions—and without any intentional "structural" cues—listeners perceive the generating key at better-than-chance levels but without high levels of agreement. The most popular key choices accounted for only slightly more than half of participants' responses. Second, the behavior of participants with absolute pitch on randomly generated melodies appears to be very similar to that of non-AP participants. Finally, to the extent that the participants do agree in their key judgments, the simple probabilistic model proposed here matches their judgments quite well—a number of alternative models and modifications to the model were tried, but none yielded better performance. We discuss each of these findings in turn.

The low level of agreement among participants in our experiment was quite surprising to us. We assume that with real tonal melodies the level of agreement would be much higher. The fact that there was such low agreement on our melodies suggests to us that listeners were often uncertain about their tonality judgments.[16] In our view, this finding casts serious doubt on the distributional view of key perception. One could argue that the melodies were too short for listeners to get an adequate "sample" of the distribution; with longer melodies, perhaps listeners would identify the generating key more reliably. But this argument seems unconvincing. For one thing, the probabilistic model was able to identify the generating key on all 60 melodies, showing that it is at least possible to identify the key in melodies of this length. Second, many tonal melodies are as short or shorter than 40 notes—for example, "Mary Had a Little Lamb" has 26 notes and "Twinkle Twinkle Little Star" has 42 notes. Listeners seem to have no trouble identifying the tonality in such melodies. Our random melodies seem to lack important cues to tonality—presumably, structural cues of some kind—that are present in real tonal music.

If listeners use other cues besides distribution to identify tonality, what might those cues be? Here, we encounter a problem mentioned earlier: Proponents of the structural view of key-finding have so far failed to advance any robust, testable key-finding model. The structural cues that have been proposed are, in themselves, hardly adequate for a key-finding algorithm. Vos's idea (1999) that an opening ascending fourth or descending fifth is an important cue fails to accommodate the many melodies that do not start with these intervals. Similarly, Butler's (1989) proposal about tritone ordering is weakened by the fact that many well-known melodies do not contain "fa-ti" tritones. ("Twinkle Twinkle Little Star" and "Mary Had A Little Lamb" are two examples of melodies that contain neither an ascending-fourth/descending-fifth opening nor any fa-ti tritones.) These proposals also do not specify how listeners distinguish parallel major and minor keys, which would share the same "fa-ti" tritone and "sol-do" ascending fourth/descending fifth. Our data also gives little support for these particular intervals as key-defining cues. We examined the 10 melodies in which the generating key was not the most popular key, to see whether these proposed structural cues explained listeners' judgments. In no case did the melody begin with an ascending fourth or descending fifth in the most popular key. In only two melodies was there a "fa-ti" tritone (as a pair of consecutive notes) in the most popular key. Thus it does not appear that these cues are essential for key identification—though they may play a small role, perhaps in nonconsecutive pitches or in combination with other structural cues.

Turning to our second conclusion, our study found no difference in key-finding performance with regard to absolute pitch. The AP, quasi-AP, and non-AP groups were almost the same in the level at which their judgments matched the generating key; they also did not significantly differ with regard to the time taken to form a key judgment on the timed experiment. None of the three AP groups—AP, quasi-AP, or non-AP—showed any bias towards white-note tonics over black-note tonics; both the AP and non-AP groups (though not the quasi-AP group) showed a slight and significant bias towards major-mode keys. Thus, AP and non-AP listeners seem very similar in their use of distributional information in key-finding.

Finally, despite the rather low level at which our participants' key judgments matched the generating keys,

[15]Table 2 also shows the subject's second most popular key for each melody. It can be seen that, in 6 of the 10 cases, the second-most popular key was the generating key.

[16]If that is the case, one might wonder why participants usually stopped the melody before it ended on the timed experiment. But this may be because they thought they were *supposed* to stop the melody before it ended.

it seems clear that they made some use of pitch-class distribution in their key judgments. The match between participants' judgments and the generating keys was much greater than chance. Thus, while pitch-class distribution does not completely determine key identification, it is at least part of the story. A question then arises as to how this distributional component of key perception can best be modeled. We first tried a very simple probabilistic model, first proposed in Temperley (2007); this model matched the participants' most popular judgments on 83.3% of our melodies. Attempts to improve the probabilistic model's performance—by ignoring a segment of the melody at the beginning or end, incorporating a bias for major-mode keys, considering scale-degree "transitions," and adding a bias towards interpreting the first note as tonic—all failed to produce any significant improvement over the original model. The probabilistic model also outperformed the Krumhansl-Schmuckler model. (We should bear in mind, however, that the K-S model was really the "original" distributional model of key-finding and the inspiration for Temperley's probabilistic model; indeed, one might well regard the probabilistic model simply as a variant of the K-S model.) Thus, in modeling the distributional component of key identification, the simple probabilistic model proposed here works remarkably well and further improvement is difficult to achieve.

So what can we conclude about the distributional view of key-finding? This is, in a sense, a question of whether the glass is half empty or half full. On the one hand, pitch-class distribution is clearly one component of key identification; and this component can be modeled quite well by a simple probabilistic model. On the other hand, the fact that only slightly more than half of our participants' key judgments matched the predictions of the distributional view suggests that there is much more to key identification than pitch-class distribution. It seems clear that structural cues of some kind—cues relating to the ordering and temporal arrangement of pitches—play a role in key perception. Current proposals, however, are hardly adequate in describing what these structural factors might be. Clearly, one of the challenges for future research will be to characterize more adequately the structural factors that affect listeners' perception of key, and to incorporate these factors into predictive, testable models.

## Author Note

We wish to acknowledge the help of a number of people on this project. Andy Flowers wrote the Java interface used in our experiments. Benjamin Anderson, Sara Ballance, Zachary Cairns, Sarana Chou, Nathan Fleshner, Jenine Lawson, and Gardiner von Trapp helped us run the experiment. And Elissa Newport provided valuable advice on experimental design and statistics. Any flaws with any aspect of the project, however, are entirely our responsibility.

*Correspondence conerning this article should be addressed to* David Temperley, Eastman School of Music, 26 Gibbs St., Rochester, NY, 14604 USA; E-MAIL: dtemperley@esm.rochester.edu

## References

AUHAGEN, W. (1994). *Experimentelle Untersuchungen zur auditiven Tonalitätbestimmung in Melodien.* Kassel: Gustav Bosse Verlag.

BHARUCHA, J. J. (1984). Anchoring effects in music: The resolution of dissonance. *Cognitive Psychology*, *16*, 485-518.

BHARUCHA, J. J., & STOECKIG, K. (1986). Reaction time and musical expectancy: Priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 403-410.

BROWNE, R. (1981). Tonal implications of the diatonic set. *In Theory Only*, *5*, 3-21.

BROWN, H. (1988). The interplay of set content and temporal context in a functional theory of tonality perception. *Music Perception*, *11*, 371-407.

BROWN, H., BUTLER, D., & JONES, M. R. (1994). Musical and temporal influences on key discovery. *Music Perception*, *11*, 371-407.

BUTLER, D. (1989). Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry. *Music Perception, 6*, 219-242.

CASTELLANO, M. A., BHARUCHA, J. J., & KRUMHANSL, C. L. (1984). Tonal hierarchies in the music of North India. *Journal of Experimental Psychology: General*, *113*, 394-412.

CHEW, E. (2002). The spiral array: An algorithm for determining key boundaries. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and artificial intelligence* (pp. 18-31). Berlin: Springer.

Cohen, A. J. (1991). Tonality and perception: Musical scales primed by excerpts from the Well-Tempered Clavier of J. S. Bach. *Psychological Research*, *53*, 305-314.

Creel, S. C., & Newport, E. L. (2002). Tonal profiles of artificial scales: Implications for music learning. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, & J. Renwick (Eds.), *Proceedings of the 7th International Conference on Music Perception and Cognition* (pp. 281-284). Sydney, Australia: Causal Productions.

Cuddy, L. L. (1997). Tonal relations. In I. Deliege & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 329-352). London: Taylor & Francis.

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*, 869-883.

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

Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception and Psychophysics*, 57, 451-62.

Deutsch, D., Henthorn, T, Marvin, E. W., & Xu, H. (2006). Absolute pitch among American and Chinese conservatory students: Prevalence differences and evidence for a speech-related critical period. *Journal of the Acoustical Society of America, 119*, 719-722.

Gregersen, P., Kowalsky, E., Kohn, N. & Marvin, E. W. (2001). Early childhood musical education and predisposition to absolute pitch: Teasing apart genes and environment. *American Journal of Medical Genetics*, *98*, 280-282.

Huron, D. (1999). *Music research using Humdrum: A user's guide.* Stanford, California: Center for Computer Assisted Research in the Humanities.

Huron, D., & Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology*, *12*, 154-171.

Janata, P., & Reisberg, D. (1988). Response-time measures as a means of exploring tonal hierarchies. *Music Perception*, 6, 161-172.

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

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*, 334-368.

Leman, M. (1995). *Music and schema theory*. Berlin: Springer.

Lerdahl, F. (2001). *Tonal pitch space*. Oxford: Oxford University Press.

Levitin, D. J., & Rogers, S. E. (2005). Absolute pitch: Perception, coding, and controversies. *Trends in Cognitive Sciences*, *9*, 26-33.

Longuet-Higgins, H. C., & Steedman, M. J. (1971). On interpreting Bach. *Machine Intelligence*, *6*, 221-241.

Marvin, E. W. (1997). Tonal/Atonal: Cognitive strategies for recognition of transposed melodies. In J. Baker, D. Beach, and J. Bernard (Eds.), *Music theory in concept and practice* (pp. 217-236). Rochester: University of Rochester Press.

Marvin, E. W., & Brinkman, A. (2000). The effect of key color and timbre on absolute-pitch recognition in musical contexts. *Music Perception*, *18*, 111-137.

Matsunaga, A., & Abe, J. (2005). Cues for key perception of a melody: Pitch set alone? *Music Perception*, *23*, 153-164.

Miyazaki, K. (1989). Absolute pitch identification: Effects of timbre and pitch region. *Music Perception*, *7*, 1-14.

Miyazaki, K. (1990). The speed of musical pitch identification by absolute-pitch possessors. *Music Perception*, 8, 177-188.

Oram, N., & Cuddy, L. L. (1995). Responsiveness of Western adults to pitch-distributional information in melodic sequences. *Psychological Research*, *57*, 103-118.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*, 74-85.

Saffran, J.R., Johnson, E. K., Aslin, R., & Newport, E. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27-52.

Schaffrath, H. (1995). *The Essen Folksong Collection* (D. Huron, Ed.). Stanford, CA: Center for Computer–Assisted Research in the Humanities.

Schmuckler, M. (1989). Expectation and music: Investigation of melodic and harmonic processes. *Music Perception*, *7*, 109-150.

Schmuckler, M., & Tomovksi, R. (2005). Perceptual tests of an algorithm for musical key-finding. *Journal of Experimental Psychology, 31,* 1124-1149.

Shmulevich, I., & Yli-Harja, O. (2000). Localized key-finding: Algorithms and applications. *Music Perception*, *17*, 65-100.

Smith, N. A., & Schmuckler, M. A. (2004). The perception of tonal structure through the differentiation and organization of pitches. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 268-286.

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin*, *113*, 345-61.

Takeuchi, A. H., & Hulse, S. H. (1991). Absolute-pitch judgments of black- and white-key pitches. *Music Perception*, *9*, 27-46.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Terhardt, E., & Seewann, M. (1983). Aural key identification and its relationship to absolute pitch. *Music Perception*, *1*, 63-83.

Toiviainen, P., & Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*, 741-766.

Vos, P. G. (1999). Key implications of ascending fourth and descending fifth openings. *Psychology of Music*, *27*, 4-18.

Vos, P. G. (2000). Tonality induction: Theoretical problems and dilemmas. *Music Perception*, *17*, 403-416.

Vos, P. G., & Van Geenen, E. W. (1996). A parallel-processing key-finding model. *Music Perception*, *14*, 185-224.

West, R. J., & Fryer, R. (1990). Ratings of suitability of probe tones as tonics after random orderings of notes of the diatonic scale. *Music Perception*, *7*, 253-258.

Yoshino, I., & Abe, J. (2004). Cognitive modeling of key interpretation in melody perception. *Japanese Psychological Research*, *46*, 283-297.