Music Perception
Fall 1996, Vol. 14, No. 1, 57–83

Using the Krumhansl and Schmuckler Key-Finding Algorithm to Quantify the Effects of Tonality in the Interpolated-Tone Pitch-Comparison Task

B. W. FRANKLAND & ANNABEL J. COHEN
Dalhousie University

We examined two models that quantified the effects of tonality on accuracy and reaction time in an intervening-tone pitch-comparison task. In each of 16 task conditions (standard tone–interpolated sequence–test tone, abbreviated as S-seq-T), the S and T tones, C_5 and/or $C\sharp_5$, were separated by a three-tone sequence that was a random arrangement of one of the four triads, c_{4Major}, c_{4minor}, $c\sharp_{4Major}$, or $c\sharp_{4minor}$. Both models were based on the tonal hierarchy (Krumhansl, 1990a; Krumhansl & Shepard, 1979) and the key-finding algorithm (Krumhansl & Schmuckler, cited in Krumhansl, 1990a); the key-finding algorithm was used to determine the best-fitting key for the first four notes of the condition (i.e., the S-seq combination). Model 1 (S-Tone Stability) determined the stability of the S tone given that key. Model 2 (T-Tone Expectancy) determined the expectancy for the T tone given that key. Over the 16 conditions, for three groups of 12 subjects, differing by level of training, mean proportion correct discrimination ranged from .53 to .95 and increased significantly across levels of musical experience. For the musically trained subjects, both models predicted performance well but neither model was dramatically more effective than the other; the combination of both models did produce an increase in predictability. For untrained subjects, tonality, as assessed by the key-finding algorithm in either model, was not significantly correlated with performance.

THE concepts of tonality and the degree of tonality pervade much research in music. By tonality we refer to the selection, arrangement, and hierarchical ordering of tones that lead to the perception of a central reference tone (the tonic), triad, or key that is typical of Western tonal music (cf. Christ, Delone, Kliewer, Rowell & Thomson, 1966; Cohen, 1991; Krumhansl, 1979; Krumhansl & Shepard, 1979; Sadie, 1980). Some research has examined the effects of tonality by contrasting the effects of

Address correspondence to B. W. Frankland, Department of Psychology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1 (e-mail: franklan@is.dal.ca) or A. J. Cohen, Department of Psychology, University of Prince Edward Island, Charlottetown, Prince Edward Island, Canada, C1A 4P3 (e-mail: acohen@upei.ca).

stimuli representing two disparate points of the tonality continuum (cf. Cohen, 1982; Cohen, Trehub, & Thorpe, 1989; Dowling, 1984; Trehub, Cohen, Thorpe, & Morrongiello, 1986). Other research has examined additional levels (e.g., three levels: Dowling, 1991; four levels: Croonen, 1994; five levels: Cuddy, Cohen, & Mewhort, 1981). These finer distinctions in degree of tonality have rested on the consensus of a small body of experts and the application of musical rules concerning harmonic progression. Although these techniques have been effective in illustrating the influence of tonality, a more objective assessment (quantification) of degree of tonality could have advantages (i.e., a basis for comparison across studies).

Assessing the degree of tonality of a musical excerpt is linked to the problem of determining the extent to which a best key can be found for the excerpt. Krumhansl and Schmuckler (Krumhansl, 1990a) offered an algorithm that determines the most appropriate key for a sequence of tones. The algorithm also provides a measure of the degree of tonality within the sequence of tones. The key-finding algorithm is supported by both theoretical and empirical studies (e.g., Krumhansl, 1979; Krumhansl & Shepard, 1979; Krumhansl & Kessler, 1982). Although it does not consider serial order effects (e.g., the effects of voice leading; cf. Brown, 1987; Brown, Butler, & Jones, 1994; Butler, 1989), the key-finding algorithm (and related hierarchy) could have pragmatic application beyond the empirical studies in which it was developed (see Butler, 1990, and Krumhansl, 1990b).

The present work used the Krumhansl and Schmuckler (Krumhansl, 1990a) key-finding algorithm (hereafter *the key-finding algorithm* and associated *tonal hierarchy*) to quantify several effects of tonality on pitch memory in an intervening-tone pitch-comparison task (hereafter *the pitch-comparison task*). The task requires subjects to make a same/different judgment about the pitch of two tones, a standard (S) followed by a test (T), separated by an intervening sequence of tones (hereafter *S-seq-T*). Krumhansl (1979) has shown the tonal relationship between the S tone, T tone, and intervening sequence affects performance. She constructed sequences that were either tonal (in a particular key) or atonal (in no particular key) and defined S tones as diatonic or nondiatonic with respect to the tonal sequences. With the tonal sequence, diatonic S tones were better remembered than the nondiatonic S tones. This was consistent with the notion that diatonic notes are more stable than nondiatonic notes within the same tonal context. With atonal sequences, the nondiatonic S tones were remembered better than diatonic S tones. This may seem puzzling: If the sequence is truly atonal, then none of the 12 chromatic notes is diatonic, and therefore, none should be preferentially remembered. After other interpretations had been eliminated (e.g., Deutsch, 1972a, 1972b, 1973, 1974, 1979), Krumhansl concluded that the atonal condition may have suggested an alternative key to which the nondiatonic tones were better matched

(Krumhansl, 1979, p. 372; see also Krumhansl, 1990a, pp. 144–147). Clearly, the ability to determine and quantify the best-fitting key for a sequence of notes would be advantageous in such a situation.

How tonality is established is of continuing interest in the field of music perception. It is well known, however, that very little information is needed to establish a best key that listeners agree on. Cohen (1991) reported that musically experienced listeners generally agreed on the key elicited by the first four note-events of a musical excerpt, the opening measures of the 12 Preludes from the *Well-Tempered Clavier* of J. S. Bach. Analysis of these excerpts revealed that they began on the tonic, contained the major triad tones, and excluded nondiatonic tones. The most favored key of each excerpt was also generally predicted by the key-finding algorithm (Krumhansl, 1990a, pp. 83–84). It is also true that particular combinations of a few notes, such as the major triad, when presented as context in a probe-tone task, lead to probe-tone profiles that are very similar to those produced by major scale contexts in which every diatonic tone is presented (cf. Brown, et al., 1994; Cuddy & Badertscher, 1987). Therefore, it was assumed, in the present study, that the presentation of a sequence of four tones (the S tone and intervening sequence) could activate a sense of key to a greater or lesser extent and that the degree of keyness of the extent to which key was activated could be quantified by the key-finding algorithm and thereby used to predict variations in performance.

The present experiment aimed to quantify further the effects of tonality, using the key-finding algorithm, in the intervening-tone pitch-comparison task. The S and T tones, C_5 and/or $C\sharp_5$, were separated by a single broken major or minor triad: $c_{4Major} = c_M$, $c_{4minor} = c_m$, $c\sharp_{4Major} = c\sharp_m$, or $c\sharp_{4minor} = c\sharp_m$. Each of the four combinations of S and T tone was presented with each intervening sequence, creating 16 different conditions (Figure 1). In order to simplify the notation, the uppercase C has been used to designate the notes used for the S and T tones (C and C♯), and the lowercase c has been used to designate the triads (c_M, c_m, $c\sharp_M$ and $c\sharp_m$) used for the sequences, which were always in the octave immediately below S and T tones.

Each of the 16 conditions consisted of a different set of tones, so each condition could elicit a different sense of tonality and tonic. Within each condition and associated sense of key, the S and the T tones would have a given stability (defined with respect to the tonal hierarchy). It can be assumed that the stabilities of the S and T tones would affect performance. Under different stimulus conditions, the S tone and T tone would have different degrees of stability: Hence, different levels of performance (accuracy) would be predicted for different conditions. For example, in line with Krumhansl's (1979) findings, S tones that are diatonic (relatively stable) within the abstracted key would be better retained than S tones that are nondiatonic (unstable) within the abstracted key.

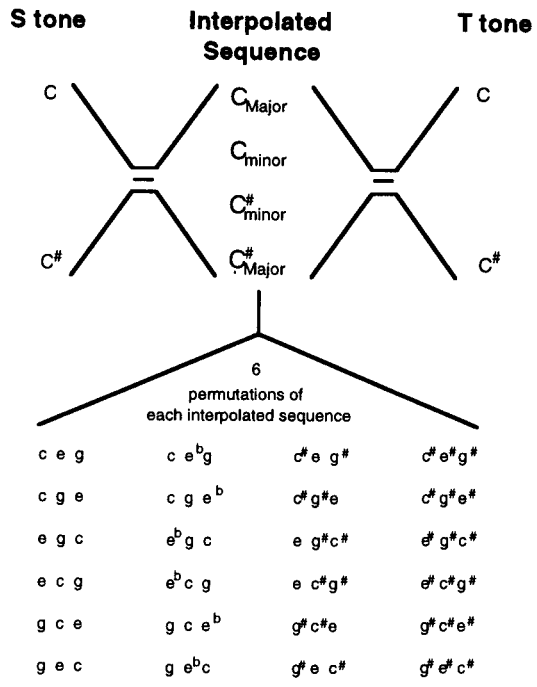

Fig. 1. The combinations of S tone, sequence, and T tone that were used in the experiment. All combinations of the two S tones, the four sequences, and the two T tones were used (16 combinations). Each sequence was presented in the six possible orders.

A range of stabilities of the S and T tones results from various relations between the S tone, sequence, and T tone (cf. Krumhansl, 1979). In the present study, the 16 different conditions of S tone, sequence, and T tone (S-seq-T) produced 14 different types of tonal relationships. There were four conditions in which both the S and T tones could be considered triad notes within the tonality implied by the sequence (i.e., C-c_M-C, C-c_m-C, C♯-$c♯_M$-C♯, C♯-$c♯_m$-C♯). However, this produced only two different types of relationships because C-c_M-C is equivalent to C♯-$c♯_M$-C♯, and C-c_m-C is equivalent to C♯-$c♯_m$-C♯. One would conjecture that, in these four conditions, the sense of key would be strong and the stability of the S and T tones high. There were two conditions in which the S tone could be considered a triad note and the T tone a diatonic note (i.e., C♯-$c♯_M$-C, C♯-$c♯_m$-C), and two conditions in which this was reversed (i.e., C-$c♯_M$-C♯, C-$c♯_m$-C♯). In these four conditions, one might expect a weaker sense of key, with high stability for triad S or T tones and lower stability for diatonic, nontriad S or T tones. There were two conditions in which the S tone could be considered a triad note and the T tone a nondiatonic note (C-c_M-C♯, C-c_m-C♯); there were two conditions in which this was reversed (C♯-c_M-C, C♯-c_m-C). In these condi-

tions, the sense of key should be weaker still, with low stabilities for the nondiatonic S or T tone. There were two conditions in which both the S and T tones could be considered diatonic notes (C-$c\sharp_M$-C & C-$c\sharp_m$-C), which should produce a fairly weak sense of tonality, and low stabilities. Finally, there were two conditions in which both the S and T tones were nondiatonic notes (C♯-c_M-C♯ and C♯-c_m-C♯). These should produce the weakest sense of key. It was predicted that the key-finding algorithm could be used to quantify such predictions.

The potential effects of tonality (as defined by Krumhansl, 1990a) on performance were quantified with two different models. Each of these models assumed that the key-finding algorithm provided the best-fitting key and that the tonal hierarchy represented a reasonable mapping of the theoretical, relative importance (stability or fit) of each chromatic note within a given key. The key-finding algorithm correlates the pattern of relative durations of each chromatic note in a sequence against the pattern of subjective ratings (derived from the probe-tone study of Krumhansl & Kessler, 1982) of the degree to which each of the 12 chromatic tones matched with each of the 24 major and minor keys. The key-finding algorithm returns the correlation coefficient (r) between the sequence and each of the 24 different major and minor keys: The highest of the 24 correlation coefficients defines the best-fitting key. Because the algorithm returns a correlation coefficient for each of the major and minor keys, there are three pieces of information that can be extracted: Which key has the highest correlation, the strength of that highest correlation, and the separation between the highest correlation and the second highest correlation (and the third highest etc.). The first corresponds to the best-fitting key. The second is a measure of the strength of the best-fitting key (i.e., whether or not the best-fitting key is well defined). The third is a measure of the ambiguity of the key abstraction. If many keys are defined with equal strength (regardless of the magnitude of that strength), then one cannot say that key has been defined (cf. the intervallic rivalry model; cf. Brown et al., 1994).

In the present experiment, for each of the 16 conditions, the highest correlation returned by the key-finding algorithm was taken to define the best-fitting key for each condition and the magnitude of that correlation was taken as a measure of the strength of the tonality implied by that condition. For computational tractability, the measure of key ambiguity was ignored. Because the correlation coefficient squared (r^2) represents the proportion of variance accounted for, the correlation coefficient squared rather than the correlation coefficient (unsquared) was used as a measure of key strength (K_s).

For both models, it was assumed that the key abstracted from the first four notes of each condition (i.e., S-seq: the S tone and the intervening sequence) would provide a frame of reference for the retention of the S

TABLE 1
Best Key, Key Strength, S-Tone Stability (Model 1), and T-Tone Expectancy (Model 2) for Each Condition

Condition	Best Key		Model 1	Model 2	
	Key	Strength[a]	S-Tone Stability[b]	T-Tone Expectancy[c]	Comp Value[d]
C - c_M - C	C_M	0.778	0.690	0.690	−1
C- c_m - C	C_m	0.847	0.687	0.687	−1
C - $c\sharp_m$ - C	$C\sharp_m$	0.538	−0.012	−0.012	−1
C - $c\sharp_M$ - C	F_m	0.736	0.290	0.290	−1
C♯ - $c\sharp_M$ - C♯	$C\sharp_M$	0.778	0.690	0.690	−1
C♯ - $c\sharp_m$ - C♯	$C\sharp_m$	0.847	0.687	0.687	−1
C♯ - c_m - C♯	C_m	0.434	−0.066	−0.066	−1
C♯ - c_M - C♯	C_M	0.349	−0.040	−0.040	−1
C - c_M - C♯	C_M	0.778	0.690	−0.089	+1
C - c_m - C♯	C_m	0.847	0.687	−0.129	+1
C - $c\sharp_m$ - C♯	$C\sharp_m$	0.538	−0.012	0.437	+1
C - $c\sharp_M$ - C♯	F_m	0.736	0.290	0.150	+1
C♯ - $c\sharp_M$ - C	$C\sharp_M$	0.778	0.690	0.034	+1
C♯ - $c\sharp_m$ - C	$C\sharp_m$	0.847	0.687	−0.019	+1
C♯ - c_m - C	C_m	0.434	−0.066	0.352	+1
C♯ - c_M - C	C_M	0.349	−0.040	0.309	+1

[a]Higher positive values indicate more precisely defined keys.
[b]A positive value indicates that the S tone does match the key created by the first four notes (i.e., the S tone and the sequence), whereas a negative value indicates that it does not.
[c]A positive value indicates that the T tone does match the key created by the first four notes, whereas a negative value indicates that it does not.
[d]Compatibility indicates that increasing discrepancy can either aid (+1) or interfere (−1) with performance.

tone and sequence. The higher the key strength, the better the frame of reference and, hence, the better the retention of the S tone. Therefore, performance would be predicted to decline as key strength declined (i.e., a positive correlation between accuracy and key strength). Table 1 presents the best-fitting key and its associated key strength for each condition.

Models 1 and 2 considered the effects of the abstracted key on the stability of the S and T tones more directly. In order to compare such effects across the 16 different conditions, two linear modifications to the tonal hierarchy were made (Figures 2 and 3). Once the best-fitting key and its associated strength had been established, it was necessary to determine the relative stability of each note given that best-fitting key. Obviously, the tonic would be the most stable and the poorest fitting nondiatonic tone would be the least stable. Therefore, the tonal hierarchy was normalized so that a value of 1.0 represented the tonic of the key and a value of 0.0

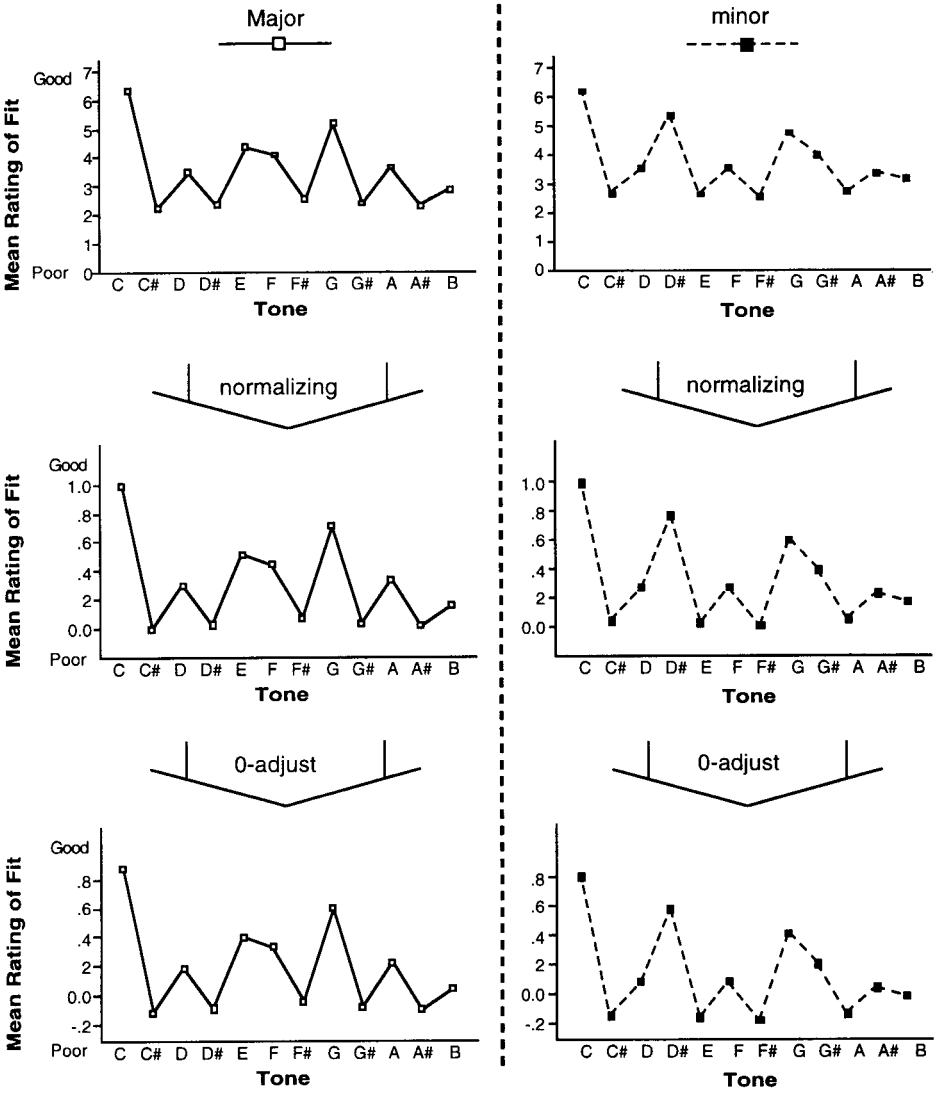

Fig. 2. The linear transformation of the tonal hierarchy. The first frame represents the original hierarchy: mean rating is the fit of each chromatic note to the keys C_{Major} and C_{minor}. The first transformation normalizes both hierarchies to the range 0.0–1.0. The second transformation adjusts the zero point so that all diatonic notes have a positive rating and all nondiatonic notes have a negative rating. Note that after the second transformation, major and minor keys do not have the same maximum and minimum values.

represented the tone with the poorest fit to the key (i.e., C♯ in the key of C). A cut line was then[1] constructed that separated the diatonic notes from the nondiatonic notes (i.e., in the key of C, the cut line would fall at the mid-

1. Because it is a *relative ranking* of the strength/stability of each S tone across the 16 conditions that is sought, the actual order of operations is largely irrelevant.

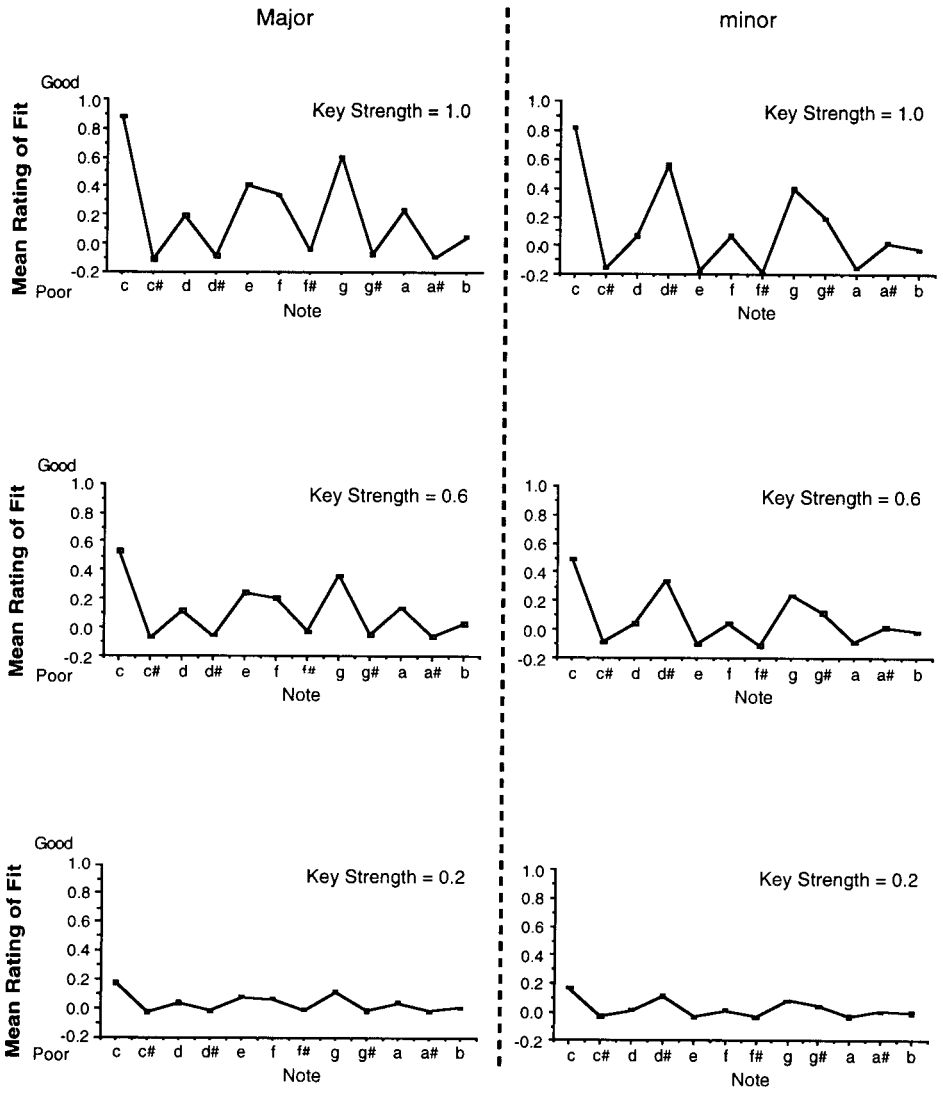

Fig. 3. The effect of key strength on the individual notes within the adjusted tonal hierarchy. In the first frame, a key strength of 1.0 is depicted (such a strength is unlikely, although not impossible, given the key-finding algorithm). The second frame depicts a key strength of 0.6. Notice that all notes are nearer to the 0 line. The third frame depicts a key strength of 0.2; all notes are very close to the zero line, even the tonic.

point between the notes B and F♯). The entire hierarchy was then shifted by a constant so that the cut line had a value of 0.0, producing the final adjusted tonal hierarchy in which diatonic notes would have a positive value and nondiatonic notes would have a negative value (Figure 2). Neither of these modifications affected the relationships between notes within a given key.

Using the best-fitting key strength, the theoretical stability of each of the 12 chromatic notes in each condition (within each abstracted key) could be determined by multiplying the stability of each note in the normalized hierarchy by the best-fitting key strength (Figure 3). Hence, there was an adjusted hierarchy for each tonality condition, and in this adjusted hierarchy, the stability of each chromatic note was determined by its role within that key (i.e., its relative importance or relative fit) in conjunction with the overall best-fitting key strength (i.e., the relative key strength of that condition). Note that when key strength is low, there is little to differentiate the diatonic tones from the nondiatonic tones.

In Model 1 (S-tone stability), the adjusted hierarchy for each condition provided a measure of the stability of the S tone within that condition (Table 1). The retention of the S tone was assumed to be related to the stability of the S tone. For example, if the S tone were a chromatic note outside the abstracted key, then the representation of the S tone would be unstable; if the S tone were a diatonic note within the abstracted key, then the representation of the S tone would be relatively stable. Accuracy was predicted to be positively correlated with stability of the S tone. A regression equation[2] could be created:

$$\text{accuracy} = a + bS_s,$$

where a and b are constants and S_s is the S-tone stability. It was assumed that subjects made their judgments immediately upon hearing the T tone, so the T tone was not included in the determination of key (and hence, the stability of the S tone) and any effects of abstracted key on the retention of the T tone were ignored.

It should be noted that Model 1 (S-tone stability) cannot account for all the differences between the 16 conditions because it cannot distinguish between complementary conditions that differ only in the T tone (e.g., C-c_m-C and C-c_m-C♯). Several previous studies (Deutsch, 1972a, 1972b, 1975) have empirically demonstrated the nonequivalence of these complementary conditions.

Model 2 (T-tone expectancy) was designed, in part, to address the weaknesses of Model 1 by considering the role of the T tone in performance. The conceptual framework defining the role of the T tone can be found in the tonal-harmonic scheme of the serial integration hypothesis of Jones (1976, 1981, 1982; see also Frankland & Cohen, 1990). It was assumed that the sense of key created by the first four tones would elicit an expectancy for further notes that fit within the abstracted key. In essence, the

2. Higher order terms (up to and including third order) were computed as a check on linearity. Unfortunately, at this point the strongest claim that can be made is that the relationship between key strength (the key-finding algorithm) and performance should be monotonic. Although the algorithm produces a ratio scale, it is at best interval, and possibly only ordinal. Except where noted, the higher order terms were nonsignificant in a hierarchical regression analysis.

expectancy for each chromatic note would be proportional to the stability of that note within the abstracted key. Therefore, using the adjusted hierarchy (as in Model 1), expected notes would have positive values and unexpected (discrepant) notes would have negative values; the actual amount of expectancy or discrepancy would be indicated by the magnitude of the value. As the amount of discrepancy increases, there is a greater instability in the recognition and encoding of the T tone, which reduces performance (cf. Model 1: S-tone stability). If the first four tones created a well-defined key, then expectancies would be well defined and strong: ranging from large positive numbers (highly expected) to large negative numbers (highly discrepant). In this case, any given T tone could be a very good match, a mediocre match, or a poor match to those expectancies. On the other hand, if the first four notes failed to create a well-defined sense of key, then the expectancies generated would be ambiguous and weak: ranging from small positive numbers to small negative numbers. In this case, the T tone could never be anything except a mediocre match to those expectancies.

The effect of such expectancies is affected by a second process (Jones, 1981, 1982). Expected events are easy to integrate (relate to previous events): Because the stimuli occur in the predicted region of tonal space, stimuli can be analyzed using a precise interval scale (in the mathematical [measurement] sense). Conversely, unexpected (i.e., discrepant) events are difficult to integrate: Because the system is not primed to evaluate events occurring in that region of pitch space, accurate encoding is difficult. However, unexpected events stand out as different. As the discrepancy between the actual and expected tone increases, there is an increasing tendency to say "different" because of the ease of detection. Such a tendency, in effect a heuristic, will improve performance when the correct answer is "different," but will impair performance when the correct answer is "same." The effect of this heuristic on performance is labeled response compatibility (Comp).

Table 1 provides the abstracted key and the expectancy (i.e., stability) of the T tone in that key (expected notes—e.g., diatonic tones—have positive values and discrepant notes—e.g., nondiatonic tones—have negative values) for each of the 16 conditions. Generally, when the correct response is "same" (the first 8 conditions), as the degree of discrepancy between the T tone and key increases, performance should decline because the encoding of the T tone is less accurate and the heuristic leads to the erroneous response "different." When the correct response is "different" (the last 8 conditions), as the degree of discrepancy increases, performance should decline because the encoding of the T tone is less stable. However, when the correct response is "different," as the degree of discrepancy increases, performance should improve because the heuristic leads to the correct response "different." To illustrate, given the S-seq combination of C-c_M, the

abstracted key is c_{Major}, and note C has a higher expectancy than the note C♯ (0.690 versus –0.089). Hence, the encoding of note C as the T tone should be more stable than the encoding of note C♯ as the T tone, and on that basis, performance should be higher for the C-c_M-C ("same") than for C-c_M-C♯ ("different"). Performance is further enhanced by the heuristic aid; because C has a strong expectancy, there is a tendency to say "same," and because C♯ has a negative expectancy (i.e., it is discrepant), there is a tendency to say "different." In the complementary S-seq condition of C♯-c_M, the abstracted key is still c_{Major}, and the note C still has a relatively higher expectancy than C♯ (0.309 versus –0.040); the encoding of C should be more stable than that of C♯ and consequently, performance should be higher for C♯-c_M-C, than for C♯-c_M-C♯. However, the heuristic interferes with accurate performance; the note C has a positive expectancy so there is a tendency to say "same," but the correct response is "different." The note C♯ has a negative expectancy, so there is a tendency to say "different," but the correct response is "same."

In this task, the effect of the T tone can be conceptualized as the joint occurrence of two independent events: The effect of expectancy on encoding and the effect of expectancy on response compatibility (i.e., tendency to say "different" when expectancy is low). Both of these events can be modeled by a linear relationship with T-tone expectancy using the equation:

$$\text{accuracy} = \underbrace{(g + hT_e)}_{\text{instability}} \times \underbrace{[\text{Comp} \times (j + kT_e)]}_{\text{compatibility}}$$

where g, h, i, j, k, and l are constants, T_e is the T-tone expectancy, and Comp is the response condition. This can be reduced to

$$\text{accuracy} = \text{Comp} \times [\, a + (b \times T_e) + (c \times T_e^2)]$$

where a, b, and c are constants, T_e is the T-tone expectancy, and Comp is the response condition. Hence, there is a single equation with one switch (Comp for compatibility). It should be noted that, as it is coded in Table 1, the Comp × a term captures any performance differences that exist between same and different conditions (note that the assignment of ±1 to these conditions is arbitrary).[3] The T-tone expectancy itself should not be a

3. Cautious interpretation of this term is advocated. Because the "different" conditions have, on average, lower T-tone expectancies (0.37 ± 0.36 for "same" versus 0.13 ± 0.21 for "different"), lower performance would be predicted in different conditions (T-tone expectancy). However, the lower average expectancies in the different conditions would also lead to the prediction of a tendency to say "different," which would aid performance. Furthermore, the Comp × a term may simply reveal a response bias in subjects (a tendency to say same or different regardless of condition), so the actual response tendency within each condition is important.

strong predictor of performance, but the Comp × b × T_e term should be. This term is the effect of T-tone expectancy corrected for condition (same/different). In effect, it is a term that represents the interaction of the T-tone expectancy with response compatibility (or heuristic).

Because each model should explain different aspects of the task, a combined model would provide a more complete account. However, given that both models are based on the same abstract construct of tonality as assessed by the key-finding algorithm, overlap between the models is inevitable. To assess this overlap, the intercorrelations between the predicted performance for both models were calculated (Table 2). The combination of Models 1 and 2 was tested using multiple regression.

Finally, given that both models are based on an abstracted cognitive structure (tonality), one might expect an effect of musical training in both models: More highly trained persons would be expected to be more sensitive to the distinctions postulated by the models and hence to show greater response variation as a function of tonality condition.

In summary, the design of this experiment was a factorial having one within-subjects factor (16 S-seq-T conditions) and one between-subjects factor (3 levels of training). As stated, each intervening sequence consisted of a major or minor triad: To control and assess sequential order effects (cf. Cuddy & Cohen, 1976; Deutsch, 1972b), each triad was presented in all of its six possible sequential orders, and the data were analyzed as a factorial having two within-subjects factors (16 S-seq-T conditions, 6 orders) and one between-subjects factor (3 levels of training). Altogether, there were 96 different condition × order combinations, with three groups of subjects. Both proportion correct and reaction time were dependent measures. The effects of tonality and order were analyzed individually within each level of training, and the ability of each model to predict performance was assessed.

TABLE 2
The Correlations Between the Predicted Scores for Both Models

	Model 1	Model 2		
	S-Tone Stability	T-Tone Expectancy	Comp	Comp × T-Tone Expectancy
Model 1: S-tone stability	1.00	0.24	0.00	−0.71[a]
Model 2: T-tone expectancy		1.00	−0.39	−0.59[b]
Compatibility			1.00	0.67[a]
Comp × T-tone expectancy				1.00

[a] $p < .01$.
[b] $p < .05$.

Method

TRIAL STRUCTURE

The stimulus set consisted of 96 five-note sequences, using an S-tone–three-note sequence–T-tone pattern. Each sequence consisted of either C_5 or $C\sharp_5$ as the S and T tones (positions 1 and 5). The intervening sequence (positions 2, 3, and 4) consisted of either the c_{Major}, c_{minor}, $c\sharp_{Major}$ or $c\sharp_{minor}$ triad, placed in the octave below the S and T tone (C_5 and $C\sharp_5$ refer to the notes of S and T tones, whereas c, e♭ (= d♯), e, f (=e♯), g, and g♯ refer to the notes of the triads). All intervening sequences were presented in all six possible permutations of order. In a pretest/familiarization phase, the stimulus set consisted of four "sequences" of two notes, using only C_5 and $C\sharp_5$.

Sequences were constructed from sine tones confined to the frequency range 523.25–1108.73 Hz (c_4–$c\sharp_5$), synthesized on a Commodore Amiga 500 computer with 8-bit resolution and a sampling frequency of 20964.4 Hz (cf. Cohen & Mieszkowski, 1989). All tones were 250 ms in duration with rise/decay times of 25 ms. The interstimulus intervals were also 250 ms. The same computer was used to provide instruction and to record responses.

PROCEDURE

Subjects sat in front of the computer in an Industrial Acoustics Company Inc. single-walled sound-attenuating room. Verbal instructions were given to each subject, and additional abbreviated instructions remained on screen for the duration of the experiment. Tones were presented monaurally to the ear of choice, at a comfortable level, through a pair of Telephonics TDH 39P headphones connected directly to the Amiga audio output.

Subjects initiated each trial by pressing the keyboard space bar and indicated their response by one of two arrow keys. The use of the left hand for the space bar and two fingers of the right hand for the arrow keys was demonstrated by the experimenter, but not enforced.

In the pretest/familiarization phase, subjects were presented with 48 trials in a single session. Each trial consisted of a pair of notes, and subjects were instructed to compare the first note with the second note and indicate, as accurately as possible, whether the two notes were the same or different in pitch. Subjects were informed that their responses were being timed, but speed was not emphasized: accuracy was. Subjects then proceeded immediately to the test phase.

In the test phase, subjects were presented with the five-note sequences and instructed to compare the first note with the last note and then to indicate, as accurately as possible, whether these two notes were the same or different in pitch. Subjects were explicitly instructed to ignore the three intervening notes of each five-note trial. Again, subjects were informed that their responses were being timed, but that accuracy, not speed, was important. Subjects completed four consecutive sessions of 96 trials, preceded by two practice trials. No trial was repeated within a session, and the order of presentation was unique and random for each session, for each subject. For the benefit of subjects, the successful recording of each response was indicated visually. Feedback (percentage correct) was given at the end of each session. Subjects were instructed to rest when they desired by not initiating the next trial, and subjects were instructed not to vocalize. The entire experiment lasted about an hour.

SUBJECTS

From the university community, three groups differing in the number of years of formal instruction on a single musical instrument were recruited. The group designated "no train-

ing" (5 men and 7 women; mean age, 24.8 ± 3.4 years) had received no musical instruction. The group designated "low training" (1 man and 11 women; mean age, 24.9 ± 6.4 years) had received 2–6 years inclusive of formal instruction (mean, 4.0 ± 1.3 years). The group designated "high training" (5 men and 7 women, mean age, 26.8 ± 9.3 years) had received eight or more years (mean, 9.5 ± 1.9 years) of formal instruction. No subject reported having absolute pitch. During the pretest/familiarization phase, subjects were further selected on the basis of two separate criteria. Subjects who did not obtain a 75% success were excluded (such errors represent both an inability to discriminate a semitone and motor response confusions while learning a speeded-response task: most subjects had errors in the first few trials). Subjects exhibiting a response bias (significantly more errors in the "same" or "different" conditions) were also eliminated. Two subjects who did not meet these criteria were replaced in order to maintain equal group sizes of 12.

Results

With each subject, each of the 96 (16 tonality conditions × 6 orders) conditions was tested four times. Performance for each subject was scored as the proportion correct, out of four, within each condition. Performance was also scored for the median reaction time, on the four trials, for each of the 96 conditions. In the reaction time analysis, those trials in which the reaction time was less than 0.0 ms (indicating an anticipatory response) were dropped. The data, presented by group, appear in Table 3.

The results for both the proportion correct and the reaction time were analyzed within parallel analyses of variance (ANOVAs) [16 (condition: S-seq-T) × 6 (orders) × 3 (level of training), the first two factors being within subjects and the last being between subjects]. The results of both analyses, over all 36 subjects, are presented in Table 4. Order was not significant for either measure, and there were no significant interactions between order and condition or training for either measure. Thus, discussion focuses on the data collapsed across different orders. Collapsing across order also helped to stabilize the measures of accuracy and reaction-time for the remaining variables: training and condition.

Training was significant for proportion correct (No, 0.65 ± 0.19; Low, 0.75 ± 0.20; High, 0.87 ± 0.18) and planned contrasts indicated a linear trend. Training had no effect on reaction time (No, 1950 ± 580; Low, 2010 ± 620; High, 1890 ± 700 ms). Condition was also significant across all 16 conditions for both proportion correct and reaction time. Generally performance declined in the less tonal, less stable, conditions. Finally, the only interaction was between training and condition for both proportion correct and reaction time.

Because of the interaction between training and condition for proportion correct, the effect of condition was examined within each level of training (i.e., a simple effects analysis of condition within training: No, Low, and High) as well as over all subjects (Overall). For proportion correct, condition was significant ($p < .01$) within each level of training: For reac-

TABLE 3
Mean Proportion Correct and Reaction Time for Each Condition for Each Group and over All Subjects

Condition	Overall	No	Low	High
		Proportion Correct		
C - c_M - C	.795 ± .195	.680 ± .203	.798 ± .158	.908 ± .168
C - c_m - C	.815 ± .185	.713 ± .190	.788 ± .178	.948 ± .093
C - $c\sharp_m$ - C	.785 ± .203	.663 ± .210	.785 ± .178	.905 ± .148
C - $c\sharp_M$ - C	.778 ± .200	.670 ± .213	.788 ± .175	.875 ± .173
C♯ - $c\sharp_M$ - C♯	.823 ± .173	.720 ± .150	.810 ± .203	.938 ± .073
C♯ - $c\sharp_m$ - C♯	.813 ± .198	.685 ± .203	.805 ± .208	.945 ± .063
C♯ - c_m - C♯	.750 ± .185	.723 ± .155	.690 ± .210	.840 ± .160
C♯ - c_M - C♯	.728 ± .198	.708 ± .175	.678 ± .220	.798 ± .195
C - c_M - C♯	.803 ± .185	.670 ± .185	.810 ± .148	.930 ± .123
C - c_m - C♯	.790 ± .185	.628 ± .168	.838 ± .138	.903 ± .135
C - $c\sharp_m$ - C♯	.738 ± .188	.658 ± .115	.713 ± .223	.845 ± .168
C - $c\sharp_M$ - C♯	.705 ± .213	.610 ± .185	.698 ± .183	.805 ± .238
C♯ - $c\sharp_M$ - C	.768 ± .218	.588 ± .190	.785 ± .190	.935 ± .113
C♯ - $c\sharp_m$ - C	.718 ± .228	.548 ± .215	.725 ± .193	.875 ± .158
C♯ - c_m - C	.640 ± .245	.553 ± .190	.625 ± .215	.748 ± .295
C♯ - c_M - C	.615 ± .245	.528 ± .215	.600 ± .215	.720 ± .283
		Reaction Time (ms x 10)		
C - c_M - C	186 ± 51	192 ± 57	194 ± 53	171 ± 43
C - c_m - C	178 ± 55	169 ± 23	183 ± 46	181 ± 84
C - $c\sharp_m$ - C	192 ± 79	183 ± 42	190 ± 50	203 ± 123
C - $c\sharp_M$ - C	185 ± 48	187 ± 51	183 ± 40	185 ± 56
C♯ - $c\sharp_M$ - C♯	182 ± 48	173 ± 34	204 ± 71	168 ± 19
C♯ - $c\sharp_m$ - C♯	185 ± 39	194 ± 47	187 ± 33	173 ± 37
C♯ - c_m - C♯	200 ± 59	180 ± 46	226 ± 73	192 ± 47
C♯ - c_M - C♯	206 ± 57	191 ± 35	226 ± 68	202 ± 62
C - c_M - C♯	183 ± 53	196 ± 63	181 ± 55	170 ± 40
C - c_m - C♯	180 ± 49	204 ± 73	167 ± 28	168 ± 27
C - $c\sharp_m$ - C♯	195 ± 55	186 ± 42	202 ± 69	196 ± 54
C - $c\sharp_M$ - C♯	201 ± 62	218 ± 81	187 ± 49	423 ± 52
C♯ - $c\sharp_M$ - C	187 ± 59	197 ± 66	199 ± 75	166 ± 22
C♯ - $c\sharp_m$ - C	201 ± 57	233 ± 76	192 ± 44	401 ± 27
C♯ - c_m - C	231 ± 92	215 ± 84	243 ± 56	460 ± 127
C♯ - c_M - C	231 ± 100	200 ± 63	252 ± 104	466 ± 125

Notes: Overall = over all subjects, No = no training group, Low = low training group, High = high training group.

tion time, condition was not significant within any level of training. Furthermore, for proportion correct, using the means for the 16 conditions, there was a high correlation (all correlations cited are Pearson's *r*) between

TABLE 4
The Source Table for Proportion Correct and Reaction Time for Three Groups of 12 Subjects (Training Level)

Source	SS	df	MS	F
		Proportion Correct		
Between Subjects	1265.76	35	36.16	
Training (Train)	461.35	2	230.68	9.46[a]
$Error_{Between}$	804.41	33	24.38	
Within Subjects	2873.34	3420	.84	
Condition (Cond)	191.59	15	12.77	16.21[a]
Order	1.36	5	.27	.35
Cond × Order	51.94	75	.69	.88
Cond × Train	54.89	30	1.83	2.32[a]
Cond × Train	9.98	10	1.00	1.27
Cond × Order × Train	93.06	150	.62	.78
$Error_{Within}$	2470.59	3135	.79	
Total	4139.10	3455		
		Reaction Time		
Between Subjects	333.03	35	9.51	
Training (Train)	8.00	2	4.00	.41
$Error_{Between}$	325.03	33	9.85	
Within Subjects	4749.19	3420	1.39	
Condition (Cond)	86.82	15	5.79	4.23[a]
Order	2.14	5	.42	.31
Cond × Order	115.31	75	1.54	1.12
Cond × Train	60.68	30	2.02	1.48[b]
Cond × Train	7.39	10	.74	.54
Cond × Order × Train	192.13	150	1.28	.94
$Error_{Within}$	4284.73	3135	1.37	
Total	5082.22	3455		

[a] $p < .001$.
[b] $p < .05$.

the High and Low Training groups, indicating the two groups performed similarly (Table 5). Neither the High nor the Low Training groups were highly correlated with the No Training group. For the reaction-time measure, the patterns of correlations between groups were similar to the pattern observed with the proportion-correct measure (Table 5). This implies that the reaction time measure serves to complement the proportion correct measure. To check this further, the correlations between reaction time and proportion correct were calculated: $r = -.96$ Overall, $r = -.71$ within the No Training group, $r = -.87$ within the Low Training group, and $r = -.78$ within the High Training group, all of which were significant ($p < .01$) Note that in general, the correlations are negative and significant, indicating no speed-accuracy trade-off (Table 5). Because the reaction time mea-

TABLE 5
The Correlation Matrices Between Groups for Proportion Correct and Reaction Time and the Correlation Matrix Within Each Group for Proportion Correct with Reaction Time

	Overall	No	Low	High
		Proportion Correct: Between Groups		
Overall	1.00	0.75[a]	0.94[a]	0.95[a]
No		1.00	0.50[b]	0.53[b]
Low			1.00	0.94[a]
High				1.00
		Reaction Time: Between Groups		
Overall	1.00	0.88[a]	0.87[a]	0.84[a]
No		1.00	0.78[a]	0.72[b]
Low			1.00	0.54[c]
High				1.00
		Proportion Correct (PC) with Reaction Time (RT): Within Groups		
PC with RT	−0.96[a]	−0.71[b]	−0.87[a]	−0.78[a]

[a] $p < .001$.
[b] $p < .01$.
[c] $p < .05$.

sure complemented the proportion correct measure while being the less sensitive (i.e., fewer significant results per group), only the proportion correct measure will be discussed further.

To compare with Model 1 (S-tone stability), a correlation was calculated between the predicted performance for the 16 conditions based on S-tone stability (see Table 1) and the observed mean proportion correct. All correlations (Table 6) except that of the No Training group were significant ($p < .01$). The results of the third-order-polynomial regression analysis (Table 6) yielded an equation for each group. For all groups and Overall, neither the quadratic nor the cubic terms were significant (computed as the incremental change in R^2 [the coefficient of multiple determination] as each higher order polynomial term was added to the equation): The linear terms were significant for the High and Low Training groups, as well as Overall, but not for the No Training group. The linear equation and its R are shown in Table 6. The lack of significant higher order terms in the equation indicates that the relationship is essentially linear.

A similar analysis was done with performance as a function of T-tone expectancy in a test of Model 2. As shown in the simple correlations of Table 7 (the top matrix), the Compatibility × T-tone expectancy was significant for all three groups: The expectancy of the T tone is a good predictor of performance only if one includes consideration of the correct response. The simple correlation of T-tone expectancy indicates that T-tone

TABLE 6

The Analysis of Model 1 (S-Tone Stability) Including, for Each Group, the Correlations Between the Model and the Observed Means

	Correlation Matrix	Regression Equation
Overall	0.65[a]	$A = (0.71 \pm 0.02) + (0.11 \pm 0.04) \times S$
No	0.12	$A = (0.64 \pm 0.02) + (0.02 \pm 0.05) \times S$
Low	0.77[b]	$A = (0.69 \pm 0.02) + (0.16 \pm 0.04) \times S$
High	0.78[b]	$A = (0.81 \pm 0.02) + (0.16 \pm 0.03) \times S$

Note that, in comparison with Table 7, correlations between the individual parameters of the model (i.e., S-tone stability) and the observed means will be the same as correlations between the model and the observed means because there is only one parameter. Second- and third-order terms are not shown because they did not account for any of the variation in any of the groups. A = predicted accuracy, S = S-tone stability (see Table 1).

[a] $p < .01$.

[b] $p < .001$.

TABLE 7

The Analysis of Model 2 (T-Tone Expectancy) Including, for Each Group, the Correlations Between the Parameters of the Model and the Observed Means, the Correlation Between the Model and the Observed Means, and the Regression Equation Using First-Order and Second-Order Terms

	Overall	No	Low	High
	Correlation Matrix			
T-tone expectancy	0.21	0.25	0.13	0.20
Compatibility	-0.54[a]	-0.79[b]	-0.31	-0.36
Compatibility × T-tone expectancy	-0.77[b]	-0.58[c]	-0.71[c]	-0.73[b]
Model (equations)	0.84[b]	0.80[b]	0.87[b]	0.82[b]

Quadratic Regression Equations for Each Group

Overall	$A = (0.75 \pm 0.01) + (-0.01 \pm 0.01) \times C + (-0.32 \pm 0.10) \times C \times T + (0.35 \pm 0.16) \times C \times T^2$
No	$A = (0.65 \pm 0.01) + (-0.05 \pm 0.02) \times C + (-0.09 \pm 0.11) \times C \times T + (0.13 \pm 0.19) \times C \times T^2$
Low	$A = (0.74 \pm 0.01) + (0.01 \pm 0.01) \times C + (-0.49 \pm 0.10) \times C \times T + (0.57 \pm 0.17) \times C \times T^2$
High	$A = (0.86 \pm 0.01) + (0.01 \pm 0.02) \times C + (-0.36 \pm 0.13) \times C \times T + (0.35 \pm 0.21) \times C \times T^2$

Third-order terms are not shown because they did not account for any of the variance in any of the groups. A = predicted accuracy; C = the term that indicates whether or not the response Compatibility affected performance; T = the T-tone expectancy (see Table 1).

[a] $p < .05$.

[b] $p < .001$.

[c] $p < .01$.

expectancy alone does not predict performance in any of the groups or Overall. However, the simple correlation for the binary coding of condition (i.e., Compatibility in Table 1) against performance indicates that Compatibility does predict performance in the No Training group ($p <$

.01). As can be seen in Table 7, the No Training group was more often incorrect for the different conditions (responding "same"), in comparison with the same conditions: This group tended to say "same." Neither the Low, nor the High Training groups evidenced such a tendency ($p > .05$). Due, no doubt, to the No Training group, the same correlation Overall was significant ($p < .05$). In the regression analysis, Compatibility was entered first, followed by Compatibility × T-tone expectancy, followed by the higher terms, up to the third order. For the No Training group, no term other than Compatibility was significant ($p < .05$). For the Low Training group, only the linear and quadratic terms were significant ($p < .01$). For the High Training group, only the linear term was significant ($p < .01$). Overall, in addition to the Compatibility term ($p < .05$), the linear term ($p < .01$) was significant. Table 7 presents the equations including all terms up to second order, as well as the R for the second-order equation. Note that the equations imply that the No Training group is distinctly different from the High and Low Training groups, which performed similarly.

In the final analyses, Models 1 and 2 were combined using multiple regression. In the first combined analysis, a stepwise approach indicated that Compatibility was the main variable for the No Training group, S-tone stability was the main variable for the Low Training group, whereas the combination of S-tone stability, Compatibility, and Compatibility × T-tone expectancy were all valuable for the High Training group. Interestingly, the important variable Overall was Compatibility × T-tone expectancy (Table 8).

To complete the exploration of the utility of the models, a hierarchical approach was taken with variables entered in order of theoretical importance: S-tone stability (the main variable of Model 1) and Compatibility × T-tone expectancy (the main variable of Model 2), followed by Compatibility alone. Table 8 presents the changes in R as variables were added to the equation. The results complemented the stepwise analysis (e.g., the equations are the same for the High Training group), but it is interesting that, for the No Training group, the stepwise equation using just the Compatibility term produced a higher value of R than did the hierarchical analysis using S-tone stability with Compatibility × T-tone. Generally, the inclusion of additional variables increased the proportion of variance accounted for over and above that which could be explained by the individual models.

Considering all results, the interpretation is as follows: For the No Training group, the important predictor is Compatibility; for the Low Training group, the important predictor is S-tone stability; and for the High Training group, the important predictors are S-tone stability with Compatibility and Compatibility × T-tone expectancy. There are other observations. The term T-tone expectancy alone was never important. In isolation, the T-tone expectancy indicates the surprise value of the T tone. However, this surprise value can either aid performance or interfere with performance; hence,

TABLE 8
The Stepwise and Hierarchical Regression Equations Predicting Performance

	Correlation Matrix			
	Overall	No	Low	High
Stepwise	0.77[a]	0.79[a]	0.77[a]	0.90[a]
Hierarchical				
S	0.65[b]	0.12	0.77[a]	0.78[a]
S and C × T	0.78[b]	0.71[b]	0.80[a]	0.82[a]
S and C × T and C	0.88[a]	0.81[b]	0.85[a]	0.90[a]

Stepwise Regression Equations for Each Group

Overall	$A = (0.74 \pm 0.01) + (-0.12 \pm 0.03) \times C \times T$
No	$A = (0.62 \pm 0.03) + (0.09 \pm 0.10) \times C$
Low	$A = (0.69 \pm 0.02) + (0.16 \pm 0.04) \times S$
High	$A = (0.77 \pm 0.02) + (0.34 \pm 0.08) \times S + (-0.08 \pm 0.03) \times C + (0.23 \pm 0.10) \times C \times T$

Hierarchical Regression Equations for Each Group

Overall	$A = (0.69 \pm 0.02) + (0.24 \pm 0.08) \times S + (0.16 \pm 0.10) \times C \times T + (-0.07 \pm 0.03) \times C$
No	$A = (0.62 \pm 0.03) + (0.09 \pm 0.10) \times S + (0.09 \pm 0.13) \times C \times T + (-0.07 \pm 0.03) \times C$
Low	$A = (0.66 \pm 0.03) + (0.29 \pm 0.10) \times S + (0.17 \pm 0.12) \times C \times T + (-0.06 \pm 0.03) \times C$
High	$A = (0.77 \pm 0.02) + (0.34 \pm 0.08) \times S + (0.23 \pm 0.10) \times C \times T + (-0.08 \pm 0.03) \times C$

For the hierarchical equations, S-tone stability was entered first, Compatibility × T-tone expectancy second, and Compatibility third. All terms are as indicated previously. A = predicted accuracy, C = the term that indicates whether or not the response Compatibility affected performance, S = the S-tone stability, T = the T-tone expectancy (see Table 1).

[a] $p < .001$.

[b] $p < .01$.

consideration of the response condition is necessary. That the term Compatibility × T-tone expectancy was significant Overall implies that tighter expectancies (e.g., longer sequences, greater constraints on the sequences) could increase the influence of expectancies in this task. The important message is that the different models do explain different aspects of performance.

Discussion

This study explored the empirical utility of the Krumhansl and Schmuckler key-finding algorithm (Krumhansl, 1990a) through two models based on the theoretical construct of the tonal hierarchy (Krumhansl, 1979, 1990a; Krumhansl & Shepard, 1979). The models examined the role of tonality within the intervening-tone pitch-comparison task (S-seq-T). The task requires subjects to make a same/different judgment about the S and T tones. The theoretical basis for the Model 1 (S-tone stability) and Model 2 (T-

tone expectancy) assumed that the first four notes of each trial would lead to the abstraction of a sense of key, that the key abstracted and its associated (key) strength could be empirically determined using the key-finding algorithm, and that the strength of the abstracted key would predict performance. In Model 1 (S-tone stability), the effect of the abstracted tonality on the retention of the S tone was examined. Here, it was assumed that the abstracted key would affect the stability of the S tone. As the stability of the S tone decreased, more errors would be produced. It was found that S-tone stability was linearly correlated with performance for the High and Low Training groups, but not for the No Training group. In Model 2 (T-tone expectancy), it was assumed that the abstracted key would create a set of tonal/harmonic expectancies (cf. Jones, 1982) that would affect the processing of the T tone. If the T tone were discrepant with expectancies, then performance would decrease. In addition to this, as the discrepancy between expectancies and the actual T tone increased, there would be an increasing tendency to say "different." Model 2 was most successful at predicting actual performance with the High Training group, implying that highly trained subjects are more capable of extracting tonal information and using it to guide subsequent processing. The Low Training group evidenced a similar trend, albeit weaker. Although it is true that Model 2 was predictive with the No Training group, this predictive ability was limited to a term that did not implicate tonality as assessed by the key-finding algorithm. It is possible that those subjects had a response bias (tendency to say "same") or it is possible that for those subjects, all S and T tones actually sounded the same. However, one should be careful about labeling all members of this group as insensitive to tonality: The average performance of the group may simply reflect the effect of averaging several different response styles. For example, in a probe-tone task, Frankland and Cohen (1990) demonstrated that the actual amount of training could be used only as a rough guide to the type (or degree) of tonality profile evidenced by any particular subject.

The reaction time analysis added further support to these findings. It was found that reaction time varied inversely with accuracy: Subjects required more time to make their decision when incorrect. This relationship held true even within the No Training group. The lack of a speed-accuracy trade-off also implies that the establishment of a tonal hierarchy is an automatic process.

Although they shared a common base, Model 1 did not overlap extensively with Model 2. The intercorrelations between the parameters used in Model 1 with those of Model 2 did not exceed $r = .71$ ($r^2 = .50$). However, although Model 2 (T-tone expectancy) demonstrated some promise, it was consistently outperformed by Model 1 (S-tone stability). In the combined analysis, Model 1 was the dominant factor: Model 2 did add some predict-

ability, particularly with the High Training group. It is likely that the stimulus construction did not favor the use of expectancies because sequences were short. However, despite such considerations, expectancy effects were demonstrated. Further work in this vein should be able to tease apart the effects of expectancy for new events from stability of old events by careful construction of longer stimuli.

This work also suggests that the key-finding algorithm can be a useful tool for designing stimuli. In particular, for this type of task, it would be possible to create different stimuli with equivalent S-tone stabilities and varying T-tone expectancies or vice versa so that the differential effects of stability and expectancy can be delineated. The use of the key-finding algorithm would also imply that dynamic structural effects (i.e., contour and/or sequential order) might be separable from more static effects (i.e., overall tonality as assessed by the algorithm). Research could benefit from a standard for comparison across studies. Butler (1989) and Brown (1987, 1988) have demonstrated that there are effects of sequential order on the abstraction of key. However, although the key-finding algorithm ignores serial order effects, it provides a good start (i.e., a useful approximation to the concept of tonality). As demonstrated here, the algorithm does have predictive ability in tasks that are distinctly different from those of its inception (i.e., the probe-tone task). It may be that the question of whether or not the algorithm works needs to be recast to address why and when the algorithm works (cf. Butler, 1989).

Although the algorithm has some predictive validity, this cannot be taken as proof of the validity of the concepts of tonal hierarchy and key abstraction. It must be remembered that it is possible for the key-finding algorithm to have some predictive validity, even if listeners do not engage in a process of key abstraction. Within the framework of the tonal hierarchy (see Krumhansl, 1990a), the key-finding algorithm captures the degree to which the note durations in the stimulus match the tonal hierarchy defining each key, which in turn, are assumed to represent the underlying stabilities (or importance) of each tone chroma once key has been established. Still within the tonal hierarchy of Krumhansl, it is arguable that listeners rest their judgments on the degree of chord cohesion (perhaps, the number of alternative keys suggested), rather than on a sense of a single tonality, as chords, keys, and the tonal hierarchy are all related (cf. Cuddy & Badertscher, 1987; Krumhansl & Kessler, 1982). It might then be argued that the algorithm, should it work, would be capturing some cognitive component related to chord cohesion. In a parallel distributed processing model (cf. Bharucha, 1987), this might correspond to the weights connecting individual note nodes to chord nodes or the amount of mutual inhibition produced when a large number of chords are activated. The algorithm might also be capturing a measure of the number of keys to which a collection of

notes (i.e., chord) belongs, with low correlations indicating a large number of keys. Within the framework of the intervallic rivalry hypothesis (i.e., Brown, 1987, 1988; Brown, et al., 1994; Butler, 1989), it is less obvious which aspects of cognition are being captured by the key-finding algorithm. Perhaps in general, it would be more fruitful to view the key-finding algorithm as providing a measure of the degree of association of the individual tones within a particular sequence. As with any mathematical modeling, it must be remembered that the key-finding algorithm is just a mathematical operation: It may work for a variety of reasons.

Nevertheless, the results generally supported the notion that assessment of the tonality of stimuli by using the key-finding algorithm can predict subjects' performance. These results are in accordance with results of previous empirical studies pertaining to the notion of a tonal hierarchy. As the degree of tonality decreases, individual tones are less stable; comparisons between individual tones become more error prone. Furthermore, given a particular degree of tonality, triadic tones are most stable, nondiatonic tones are least stable, with the remaining diatonic notes falling in between. The combination is similar to the concept of contextually mediated harmonic stability described by Bharucha and Krumhansl (1983). They found that the perceptual distance between two chords systematically changed as a function of the explicit tonal context. In the present work, the ability to determine the distance between two notes systematically varied as a function of the implicit tonal context. The present work is similar to that of Krumhansl (1979), which demonstrated asymmetric similarity ratings for tone pairs within a tonal context. Nondiatonic tones were judged more similar to diatonic tones than the reverse. This implied a tendency for less stable tones to move toward more stable tones and the associated tonal center (Krumhansl, 1979, p. 34). The present work (Model 1) demonstrated that even those tones defining the context are susceptible to this motion. In a similar vein, the present work (Model 2) supported the notion of tonal/harmonic expectancy schemes (Jones, 1982). Schmuckler (1989) has demonstrated that subjects create expectancies that can be related to the defined tonal structure of the piece. Unyk and Carlsen (1987) have demonstrated that violations of strong expectancies produce more errors. In the present work (Model 2), expectancies were defined by the abstracted key and the strength of that expectancy was correlated with performance (when response condition was included as a factor). Bharucha and Stoeckig (1987) demonstrated similar effects with chords: Subjects heard two chords in succession and were asked to label the second as either major or minor. Errors and reaction times increased as a function of the tonal expectancy.

As demonstrated here, consideration of two aspects of the procedure is necessary for accurate application: the separation of the choice of best key (the highest correlation returned by the algorithm) from the strength asso-

ciated with that best key (the magnitude of the correlation, or correlation squared). Properly, one would also need to consider the question of key ambiguity, an issue not addressed in this work. The magnitude of the correlation coefficient of the individual sequence with the best-fitting key is a measure of how well the set of tones matches their best-fitting key. As such, the correlation can be taken as a measure of how related the individual tones in the sequence are to each other if/when the listener abstracts that best-fitting key. This measure of association reflects an upper limit for the value: If the listener does not extract the best-fitting key, then the individual tones would have an even lower association. This has consequences for those researchers who are concerned with other effects, such as sequential order. It is possible that the ambiguity evident in the literature about various aspects of serial order is due, in part, to the confounding effects of variations in level of tonality. By using the correlation coefficient squared as a measure of the degree of tonality (as opposed to the key) or more simply, the degree of internote association, one could covary out (using analysis of covariance [ANCOVA], or possibly randomized-blocks ANOVA) any effects of note set when examining the specific effects of serial order (cf. Dowling, 1991; Frankland & Cohen, 1990). More elaborate analysis might be able to separate the sequential position effects of an individual note from its more static contribution to the tonality or key of a sequence.

Over levels of training, the effects followed predictions such that there were more effects of tonality (i.e., as assessed by the key-finding algorithm) for groups with some training (2 or more years of formal training) and minimal effects for groups without any formal training. This agrees with Krumhansl and Shepard (1979) and Frankland and Cohen (1990), who found, in a probe-tone task, that individuals could be classified by their "tonality" profile. The groups of subjects that showed essentially no sense of key evidenced a much lower level of training. Consistently, in combination (Models 1 and 2 together) and individually (each model individually) the regression equations showed the similarity of the High and Low Training groups; the main difference between these groups was found in the intercepts, indicating that overall, the High Training group performed better. The No Training group differed from the High and Low Training groups in both slope and intercept. Hence, one would conclude that any training (recall that two years of formal training was the criterion for the Low Training group) will create the same tonality structure (as assessed by the key-finding algorithm; Krumhansl, 1990a). Beyond two years, further training simply enhances performance in this task.

Given that the observed results did not follow predictions exactly for any one model, it is necessary to ascertain whether another model might better account for the results obtained. Including all five notes (i.e., the S tone, the sequence, and the T tone) in the determination of key did not

alter the findings significantly. Several alternatives were considered, but none performed better than the previous, and all were more complex. Therefore, as a guide for future research, the essential aspects will be presented briefly.

The first alternative was a simple extension of the Model 1. Here, all five notes (S tone, sequence, and T tone) were included in the determination of the best-fitting key, and consequently, the measure of S-tone stability. This did not substantially alter the correlations between the model and performance. The range of correlations, within each group and overall, ranged from .17 to .67. A second alternative expanded on Model 1, which was based on the stability of the S tone. If the S tone is unstable, or even if the S tone is not perfectly stable, then the memory trace of the S tone may drift off its true frequency. It was assumed that this drift would be, at most, to a nearest neighbor in pitch space. In other words, the drift would be limited to a semitone. Furthermore, this drift would have a preferential direction: An unstable tone would drift to a more stable neighbor and not to an unstable neighbor. The Krumhansl hierarchy was used to predict the amount and direction of the drift. However, the computational difficulty of the model (e.g., the number of parameters; that preferential drift can result in the correct response for the wrong reasons) resulted in a more complex model that correlated with actual data only to the same degree as the previously presented models (r = .08 to .79). There is a third alternative that must be considered. If one considers all five tones of the trial (i.e., the S-seq-T combination), then, for each and every trial, it is possible to find a single diatonic set (i.e., key) that contains those five tones. That means that none of the tones can be considered out-of-key (i.e., nondiatonic). The listener can then base judgments on the dimension of in-chord or out-of-chord. In fact, the listener can make such judgments without ever explicitly defining a best-fitting key, because many chords (i.e., those represented by the three, four, or five notes of each trial) belong to a number of diatonic sets. Although the theoretical rationale is different, this interpretation is similar to that of the first alternative suggested.

In conclusion, this work has shown that the Krumhansl and coworkers (Krumhansl, 1990a; Krumhansl & Shepard, 1979) hierarchy and, more specifically, the Krumhansl and Schmuckler (Krumhansl, 1990a) key-finding algorithm can be used to quantify, and predict, the effects of tonality within the intervening-tones pitch-comparison task. As such, the key-finding algorithm may be useful as a measure of the internote association for any given sequence. However, because the algorithm was successful only for those listeners with some training, the application of the algorithm is restricted to those with some training. It is implied that individuals automatically establish a tonal hierarchy and its strength affects the retention and/or encoding of information. Finally, the results are consistent with the

view that the ability to abstract tonality benefits from initial musical training. Alternatively, one could consider an individual's profile as a measure of that individual's tonal hierarchy, with the realization that training tends to lead to similar profiles.[4]

References

Bharucha, J. (1987). MUSACT: A connectionist model of musical harmony. *Proceedings of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Bharucha, J. & Krumhansl, C. L. (1983). The representation of harmonic structure in music: Hierarchies as a function of context. *Cognition, 13.* 63–102.

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

Brown, H. (1987). Tonal hierarchies and perceptual context: An experimental study of music behaviour. *Psychomusicology, 7,* 77–90.

Brown, H. (1988). The interplay of set content and temporal context in a functional theory of tonal perception. *Music Perception, 5,* 219–250.

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.

Butler, D. (1990). Response to Carol Krumhansl. *Music Perception, 7,* 325–338.

Christ, W., DeLone, R., Kliewer, V., Rowell, L. & Thomson, W. (1966). *Materials and structure of music.* Englewood Cliffs, NJ: Prentice-Hall.

Cohen, A. J. (1982). Exploring the sensitivity to structure in music. *Canadian University Music Review, 3,* 15–30.

Cohen, A. J. (1991). Tonality and perception: Musical scales primed by excerpts from the *Well Tempered Clavier* of J. S. Bach. *Psychological Research, 28 ,* 255–270.

Cohen, A. J., & Mieszkowski M. (1989). Frequency synthesis with the Commodore Amiga for research on perception and memory of pitch. *Behavior Research Methods Instruments & Computers, 21,* 623–626.

Cohen, A. J., Trehub, S. E. & Thorpe L. A., (1989). Effects of training and uncertainty on melodic information processing. *Perception & Psychophysics, 46,* 18–36.

Croonen, W. L. M. (1994). Effects of length, tonal structure, and contour in the recognition of tone series. *Perception & Psychophysics, 55,* 623–632.

Cuddy, L. L., & Badertscher, B. (1987). Recovery of the tonal hierarchy: Some comparisons across age and levels of musical experience. *Perception & Psychophysics, 41,* 609–620.

Cuddy, L. L., & Cohen, A. J. (1976). Recognition of transposed melodic sequences. *Quarterly Journal of Experimental Psychology, 28,* 255–270.

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

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

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

4. This research was based on data collected for the Masters of Science thesis by B. W. Frankland (1990) and was supported by a grant from the Natural Science and Engineering Research Council of Canada to A. J. Cohen.

Deutsch, D. (1973). Octave generalization of specific interference effects in memory for tonal pitch. *Perception & Psychophysics, 13,* 271–275.
Deutsch, D. (1974). Generality of interference by tonal stimuli in recognition memory for pitch. *Quarterly Journal of Experimental Psychology, 26,* 229–234.
Deutsch, D. (1975). Facilitation by repetition in recognition memory for tonal pitch. *Memory & Cognition, 3,* 263–266.
Deutsch, D. (1979). Octave generalization and the consolidation of melodic information. *Canadian Journal of Psychology, 33,* 201–205.
Dowling, W. J. (1984). Musical experience and tonal scales in the recognition of octave scrambled melodies. *Psychomusicology, 4,* 13–32.
Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception & Psychophysics, 50,* 305–313.
Frankland, B., & Cohen, A. J. (1990). Expectancy profiles generated by major scales: Group differences in ratings and reaction time. *Psychomusicology, 9,* 173–192.
Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83,* 323–355.
Jones, M. R. (1981). Music as a stimulus for psychological motion: Part I. Some determinants of expectancies. *Psychomusicology, 1,* 34–51.
Jones, M. R. (1982). Music as a stimulus for psychological motion: Part II. An expectancy model. *Psychomusicology, 2,* 1–13.
Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11,* 346–374.
Krumhansl, C. L. (1990a). *Cognitive foundations of musical pitch.* New York: Oxford University Press.
Krumhansl, C. L. (1990b). Tonal hierarchies and rare intervals in music cognition. *Music Perception, 7,* 309–324.
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.
Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 579–594.
Sadie, S. (Ed.) (1980). *The new Grove dictionary of music and musicians.* London: Macmillan Publishers Ltd.
Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception, 7,* 109–149.
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 & Performance, 12,* 295–301.
Unyk, A., & Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology, 7,* 3–23.