



Expectancy in melody: tests of the implication-realization model

E. Glenn Schellenberg*

Department of Psychology, University of Windsor, Windsor, Ontario, Canada N9B 3P4

Received April 25, 1994; final version accepted January 6, 1995

Abstract

The implication-realization model's description of tone-to-tone expectancies for continuations of melodies was examined. The model's predictions for expectancies are described with a small number of principles specified precisely in terms of interval size and direction of pitch. These principles were quantified and used to predict the data from three experiments in which listeners were required to judge how well individual test tones continued melodic fragments. The model successfully predicted listeners' judgments across different musical styles (British and Chinese folk songs and Webern *Lieder*), regardless of the extent of listeners' musical training (Experiments 1 and 2) or whether they were born and raised in China or the U.S.A. (Experiment 3). For each experiment, however, the collinearity of the model's predictors indicated that a simplified version of the model might predict the data equally well. Indeed, a revised and simplified model did not result in a loss of predictive power for any of the three experiments. Convergent evidence was provided in a reanalysis of data reported by Carlsen (1981) and Unyk and Carlsen (1987), whose listeners were required to sing continuations to two-tone stimuli. Thus, these findings indicate that the implication-realization model is over-specified. The consistency that was found across experimental tasks, musical styles, and listeners raises the possibility, however, that the revised version of the model may withstand the original model's claims of universality.

1. Expectancy in melody: tests of the implication-realization model

Because music is found in all known cultures, it may well embody fundamental psychological principles of perception, thought, and action.

* E-mail: schelle@uwindsor.ca.

Indeed, psychological constraints may underlie cross-cultural musical universals (or near-universals) such as: (1) octave equivalence, (2) logarithmic pitch scales, (3) five to seven discrete pitches per octave, (4) hierarchies of tonal stability, (5) melodic contour as an organizing device, and (6) a beat framework for rhythmic organization (Dowling & Harwood, 1986). One approach to the psychological study of music attempts to apply principles of perceptual organization in audition to music (e.g., Bregman, 1990, Ch. 5). Another approach focuses on learned aspects of music (e.g., Krumhansl, 1990).

A synthesis of the disparate approaches is available in Narmour's (1990, 1992) theory of melody, which is called the *implication-realization* (I-R) model. Despite its origin in music theory, the model has potential relevance to auditory pattern processing in general and to music perception in particular. According to Narmour, the perception of melody is a function of a small number of universal principles that act in conjunction with style-specific factors. The present paper provides an empirical test of this claim.

Narmour's (1990, 1992) model, with its emphasis on the psychological basis of musical structure, is in the theoretical tradition of Meyer (1956, 1973). Both theorists emphasize the foreground, or tone-to-tone level of music, and characterize the *set* of possible continuations (rather than a single continuation) suggested by an incomplete musical pattern as *implications* rather than *expectations*. There are notable differences, however. Whereas Meyer (1973) does not attempt to disentangle general psychological principles from products of learning, Narmour (1990, 1992) clearly distinguishes between universals and style-specific norms in the cognition of melodies. According to the I-R model, humans have a "genetic code" or inborn set of principles that are operative when listening to melodies (Narmour, 1989). The theory is psychologically relevant not only because of its content, but also because of its precise specification of the principles which allows for verification by empirical means.

2. The implication-realization model

The I-R model describes the cognition of melodies as a series of closures, implications, and realizations. A number of factors are proposed to contribute to a sense of closure (rest, or release from tension). Closure occurs when two successive tones have the following properties: (1) the second tone is longer than the first, (2) the second tone occurs on a stronger beat than the first, or (3) the second tone is more stable in the established key or mode than the first. Closure also occurs when three successive tones create a large interval followed by a smaller interval, or when they change pitch contour (up-to-down, up-to-lateral, down-to-up, down-to-lateral, lateral-to-up, and lateral-to-down: "lateral" being a repeated tone). Each factor can occur alone or in combination with one or more of the others.

Closure, then, is a matter of degree, depending on the number of contributing factors.

An interval that is unclosed by the aforementioned criteria generates implications for listeners: the Gestalt principles of proximity, similarity, and symmetry (see Koffka, 1935; Kohler, 1947) are said to contribute to these implications (Narmour, 1990, 1992). Because an unclosed interval generates implications for the continuation of a melody, it is called an *implicative* interval. The next interval (formed by the second tone of the implicative interval and the immediately following tone) is called a *realized* interval. The realized interval need not conform to melodic implications. Indeed, violations of implications produce particular affective and aesthetic effects (Narmour 1990, 1992). Five principles describe the core melodic implications of the I-R model: *registral direction*, *intervallic difference*, *registral return*, *proximity*, and *closure* (defined below). These principles are articulated in terms of pitch direction (upward, downward, or lateral) and interval size (the distance in pitch between two tones), which are considered the primary parameters of melody. Although Narmour (1990, 1992) uses two of the five principles (registral direction and intervallic difference) to classify all possible combinations of implicative and realized intervals into 12 mutually exclusive categories, called *basic melodic structures*, all five principles are considered in the present report. It should also be noted that although Narmour (1990, 1992) does not explicitly describe the I-R model as a combination of these five factors, he considers the present interpretation to be a fair representation of his model (E. Narmour, personal communications, February, 1990; April, 1990; May, 1991; June, 1991).

The grid in Fig. 1 is useful for describing the principles. The vertical axis corresponds to the implicative interval, ranging from 0 to 11 semitones, and is subdivided into small and large intervals. Narmour (1990, 1992) defines small implicative intervals as five semitones or smaller and large implicative intervals as seven semitones or larger. Because he considers implicative intervals of six semitones (tritones) to be a threshold, functioning as either small or large depending on the context, such intervals were excluded from the grid (Fig. 1). Implicative intervals of 12 semitones (octaves) were also excluded because of Narmour's view that octave equivalence makes them atypical examples of large intervals. The horizontal axis of the grid in Fig. 1 corresponds to the realized interval, ranging from 12 semitones in the opposite direction of the implicative interval to 12 semitones in the same direction of the implicative interval. (In principle, the figure could be extended indefinitely in both horizontal directions and also downward, but intervals larger than these are infrequent in most musical styles.) No distinction is made between ascending and descending intervals because the basic principles of the I-R model apply to both directions.

To test the claims of the I-R model, a quantitative predictor variable was constructed for each of the model's five principles. The first principle, *registral direction*, concerns the pattern of increasing and decreasing pitch

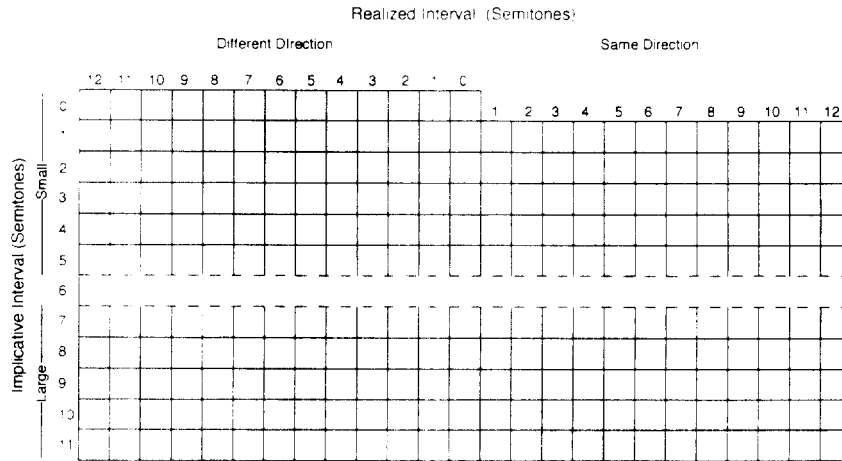


Fig. 1. This grid is useful for describing the I-R model's (Narmour, 1990, 1992) principles governing melodic expectancy. An implicative interval is an unclosed interval that creates expectancies for continuation. The vertical axis corresponds to the size of the implicative interval, ranging from 0 semitones to 11 semitones, subdivided into small (0–5 semitones) and large (7–11 semitones) implicative intervals. A realized interval follows an implicative interval (formed by the second tone of the implicative interval and the next tone). The horizontal axis corresponds to the size of the realized interval, ranging from 12 semitones in a direction different to that of the implicative interval to 12 semitones in the same direction as the implicative interval.

(melodic contour). It states that small intervals imply melodic continuation in the same direction, whereas large intervals imply a reversal of direction. As shown in the top panel of Fig. 2, combinations of implicative intervals and realized intervals that satisfy this principle (e.g., C_4 - D_4 - E_4 or C_4 - A_4 - G_4)¹ were coded as 1; cases not satisfying the principle (e.g., C_4 - D_4 - B_3 or C_4 - A_4 - B_4) were coded as 0. The dummy (all-or-none) variable created in this way was called *REGISTRAL DIRECTION*.

The second principle, *intervallic difference*, concerns the relative sizes of implicative and realized intervals. It states that small intervals imply similarly-sized intervals whereas large intervals imply smaller intervals. Narmour's (1990, 1992) definition of a similarly-sized interval depends on whether registral direction stays the same or changes. In the former case, similarly-sized means the same size plus or minus three semitones; in the latter case, similarly-sized means the same size plus or minus two semitones. As shown in the second panel of Fig. 2, a dummy predictor variable, *INTERVALLIC DIFFERENCE*, was formed by coding combinations of implicative and realized intervals that satisfied this principle (e.g., C_4 - D_4 - E_4 or C_4 - A_4 -

¹The subscript denotes the octave from which the tone is drawn. Tones have the same subscript as the closest C tone that is lower in pitch. C_4 is Middle C.

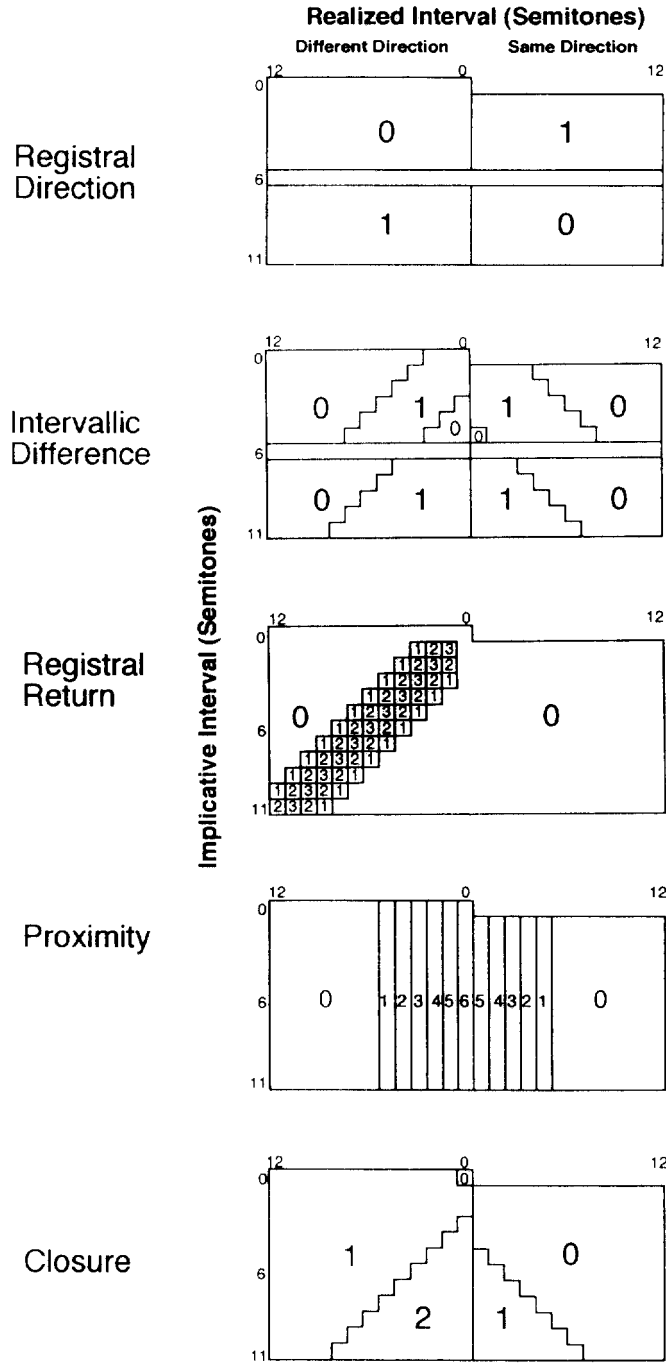


Fig. 2. The figures show how the five principles of the I-R model were quantified in order to test the model.

B₄) as 1; all other combinations (e.g., C₄-D₄-A₄ or C₄-A₄-E₅) were coded as 0.

The third principle, *registral return*, refers to cases in which the second tone of the realized interval reverses pitch direction (upward–downward or downward–upward) and is within two semitones of the first tone of the implicative interval. Thus, registral return refers to patterns that are symmetrical (ABA) or approximately symmetrical (ABA'). According to the theory, listeners recognize this pattern as a melodic archetype. This principle is graded; patterns become less archetypal as they deviate from exact symmetry. Cases of exact symmetry (e.g., C₄-F₄-C₄) were coded as 3; cases in which the second tone of the realized interval was 1, 2, or 3 or more semitones away from the first tone of the implicative interval (e.g., C₄-F₄-C#₄, C₄-F₄-D₄, C₄-F₄-E₄, respectively) were coded as 2, 1, and 0, respectively (see Fig. 2, third panel). This predictor variable was called REGISTRAL RETURN.

The fourth principle, *proximity*,² describes a general preference for small realized intervals, consistent with the cross-cultural prevalence of small intervals in music (Dowling & Harwood, 1986) and with empirical findings of perceptual grouping based on pitch proximity (e.g., Bregman, 1990; Deutsch, 1978; Handel, 1989). In particular, Narmour (personal communication, June, 1991) considers realized intervals that are small (five semitones or less) to embody proximity. This principle is also assumed to be graded; the greater the degree of proximity, the greater the implication. Accordingly, the predictor variable PROXIMITY was coded as 6 for realized intervals of 0 semitones (e.g., C₄-G₄-G₄), 5 for realized intervals of 1 semitone (e.g., C₄-G₄-F#₄), and 4, 3, 2, and 1 for intervals of 2 (e.g., C₄-G₄-F₄), 3 (e.g., C₄-G₄-E₄), 4 (e.g., C₄-G₄-D#₄), and 5 (e.g., C₄-G₄-D₄) semitones, respectively, as shown in Fig. 2 (fourth panel). For realized intervals larger than 5 semitones (e.g., C₄-G₄-C₄), PROXIMITY was coded as 0.

The fifth and final basic principle of the I-R model is *closure*. As noted, two circumstances involving pitch direction and interval size contribute to closure (i.e., when the pitch contour reverses direction, and when a large interval is followed by a smaller interval). These circumstances generate three levels of closure corresponding to both, one, or neither circumstance holding (e.g., C₄-G₄-F₄, C₄-G₄-A₄, C₄-G₄-D₅, respectively); the predictor variable CLOSURE was coded as 2, 1, or 0, respectively, as shown in Fig. 2 (bottom panel). This system of classification is independent of other factors producing closure (duration, meter, and harmony), which were held constant in the present experiments (either by the design or by statistical methods).

² Although Narmour (1990, 1992) describes in detail the influence of proximity on melodic expectancy, he does not explicitly define proximity *per se*. The present definition of proximity is derived from Narmour's definition of small intervals (five semitones or smaller) and assumes that the component tones of small intervals are proximate.

In sum, five principles (registral direction, intervallic difference, registral return, proximity, and closure) constitute the core of the I-R model's description of the implications produced by unclosed intervals. Recall, however, that Narmour (1990, 1992) considers these principles to operate in conjunction with style-specific factors. Accordingly, a quantitative covariate coding style-specific tonal influences in the musical stimuli was included in the analyses to control for their effect on listeners' responses (except in Experiment 2, which used atonal stimulus materials).

The I-R model is of considerable psychological interest for a number of reasons. The precise, quantifiable principles are based on perceptual processes known to operate in audition and vision. Moreover, the model is presumed to apply to music listening in different cultures and historical periods, regardless of the listener's musical training and experience. Finally, although the model's principles are assumed to be universal, they are hypothesized to operate in conjunction with principles learned from specific musical styles.

In previous research on musical expectancy, Carlsen (1981) and Unyk and Carlsen (1987) presented two tones, representing the beginning of a melody, and instructed listeners to sing the continuation of the melody (had it not been interrupted). The investigators transcribed the resulting vocal productions, and analyzed the tones immediately following the two-tone stimulus intervals, which consisted of all possible ascending and descending intervals smaller than or equal to 12 semitones (a total of 25 stimulus intervals). These intervals produced distinctive *expectancy profiles*, or distributions of response-tone frequencies. The results revealed a number of trends consistent with Narmour's (1990, 1992) account. For example, all stimulus intervals resulted in multiple response tones, implying that listeners expect a set of tones (rather than a single tone), and that these expectancies differ across listeners. Moreover, listeners tended to produce response tones proximate in pitch to the final tone of the stimulus interval. Finally, response tones frequently returned to the starting pitch of the stimulus interval, creating a symmetrical pattern. These qualitative observations will be more thoroughly examined below using the quantified principles of the I-R model.

3. Overview of the experiments

The experiments reported here examined the psychological validity of the I-R model's description of melodic expectancies. These experiments involved three musical styles, listeners varying in musical training, and two experimental measures. In Experiment 1, the applicability of the model's principles to a familiar (Western) style was tested among listeners who differed in the extent of their musical training. The relevance of the principles to atonal music (an unfamiliar but Western style) and to non-

Western (Chinese) music was evaluated in Experiments 2 and 3, respectively. The dependent measures in the three experiments were judgments of the appropriateness of possible continuation tones (i.e., different test tones following an implicative interval in a melody). Multiple regression analyses were used to determine how well listeners' judgments conformed to the predictions of the model. Finally, the melodic production data of Carlsen (1981) and Unyk and Carlsen (1987) were reanalyzed to determine how well the I-R model could predict results obtained with an alternative method.

The null hypothesis in each of the present experiments was that the I-R model would predict listeners' responses no better than one would expect on the basis of chance alone. This test of the model may seem relatively weak, but there are no alternative models to use for comparative purposes. Other models of musical expectancy based on principles of proximity and good continuation (Deutsch & Feroe, 1981; Simon & Sumner, 1968) assume that listeners use pattern induction processes to develop expectancies for successive events in melodies. The formal representations consist of rules (such as *same* and *next*) and alphabets (such as the chromatic scale, diatonic scales, or triads). Operators are applied recursively (with different operations and alphabets at different levels) to produce ordered hierarchical structures. Such coding models, although applicable to selected musical excerpts, are limited in their ability to describe even small segments of most musical pieces. Indeed, empirical support for this approach comes exclusively from studies using musical fragments artificially constructed to conform to these models (e.g., Boltz, Marshburn, Jones, & Johnson, 1985; Deutsch, 1980). Moreover, Boltz and Jones (1986) have shown that operators at higher hierarchical levels may not be psychologically relevant.

In each of the present experiments, the initial evaluation of the I-R model considered how well it predicted the data above and beyond chance levels. A subsequent evaluation compared a proposed revision of the model with the original model as elaborated by Narmour (1990, 1992).

EXPERIMENT 1: JUDGMENTS OF WESTERN TONAL MELODIES

Experiment 1 was designed to test the I-R model (Narmour, 1990, 1992) with melodic materials drawn from Western tonal music. Tonal music refers to music in which one tone functions as a reference point or tonic (i.e., the tone that gives its name to the scale; C is the tonic of the C major scale). Western major and minor scales, called diatonic scales, use a subset of seven tones from the chromatic scale, which divides an octave (a 2:1 ratio of fundamental frequencies) into 12 equidistant steps (semitones) on a logarithmic pitch scale. Adjacent tones in a major scale are either one or two semitones apart. A major scale is formed by starting at any position in the chromatic scale and including tones successively 2, 2, 1, 2, 2, 2, and 1

semitones higher. The eighth, or final, tone is an octave above the starting tone; the starting and final tones are both referred to as the tonic. One form of minor scale, the *natural* minor, is identical to the major scale but starts three semitones below the tonic of the relative major key; thus, successive intervals in this scale are 2, 1, 2, 2, 1, 2, and 2 semitones. In music theory, the tonic is considered the most stable tone in both major and minor scales, followed by the other tones in the tonic triad of the scale (the three-tone chord containing the tonic as well as the third and fifth degrees of the scale), the remaining tones in the scale, and finally by tones outside the scale.

The differences in stability described by music theory have demonstrable consequences for music perception and memory (see Dowling & Harwood, 1986; Handel, 1989; Krumhansl, 1990, 1991 for summaries). Accordingly, a covariate based on the hierarchy of stability values measured by Krumhansl and Kessler (1982, reported in Krumhansl, 1990) was included in the model to account for differences in stability among test tones for a given melody. The numerical values of this “tonal hierarchy” correspond qualitatively to predictions from music theory (the tonic of the key has the highest stability value, followed by the other tones in the tonic triad, other tones in the scale of the key, and, finally, tones outside the scale). Thus, if a test tone was the tonic of the key, it was assigned the highest numerical value from the tonal hierarchy. Test tones that were the fifth and third scale degrees had the next highest values; other test tones had relatively low values. The tonality covariate controlled for learned tonal influences on melodic expectancies by partialling out variation in ratings due to differences in the stability of test tones in the established key of each stimulus melody. Because this factor is considered to result from exposure to music, the covariate might be relatively more important for listeners with more musical training. Alternatively, because all of the listeners in the present experiment had a lifetime of exposure to Western music, the extent of their implicit knowledge of Western scale structure might be similar.

The melodic materials in the present experiment were British folk songs collected by Cecil Sharp (1920) and Ralph Vaughan Williams (Palmer, 1983). These songs were chosen because of their stylistic consistency and moderate familiarity to listeners from the United States. Songs from both major and minor keys were included so that the results would not be specific to either mode. British folk songs in minor keys tend to use the natural minor scale, which contains the same intervals as the relative major scale, but has a different tonic or starting point. Such folk songs are also traditionally sung solo without instrumental accompaniment, providing an appropriate context for testing the I-R model of melody.

Melodic fragments (shown in musical notation in Fig. 3) consisted of approximately one and a half phrases, chosen such that the final interval in each fragment was unclosed and therefore implicative. Two fragments (one ending in an ascending interval, the other in a descending interval) were selected for each interval tested. The small implicative intervals were 2

Fragment 1

Fragment 2

Fragment 3

Fragment 4

Fragment 5

Fragment 6

Fragment 7

Fragment 8

Fig. 3. The figure shows the melodic fragments used in Experiment 1. The fragments are taken from British folk songs.

semitones and 3 semitones, selected to be representative of Narmour's "small" intervals (i.e., 1–5 semitones); the large intervals were 9 semitones and 10 semitones, selected to be representative of "large" intervals (i.e., 7–11 semitones). Listeners rated how well individual test tones added to the ends of the fragments continued these fragments. Test tones consisted of all diatonic tones (i.e., tones from the scale of the key of the fragment) within an octave from the last tone of each fragment; non-diatonic tones were excluded to eliminate the possibility that listeners might make their ratings primarily on the basis of key membership. The rating data were tested against the quantitative formulation of the model. The tonal hierarchy (Krumhansl & Kessler, 1982) was included as a covariate, denoted TONALITY, to control for effects of familiarity with the tonal structure of the stimuli.

4. Method

4.1. Participants

The listeners were 20 members of the Cornell University community. Individual listeners were classified into one of two broad categories according to their musical training. Those with *limited training* ($n = 10$) had no musical involvement during the past 2 years and less than 6 years of musical training or regular playing experience. Listeners with *moderate training* ($n = 10$) had regularly played or studied music during the previous 2 years and had at least 6 years of musical training or regular experience playing music. Thus, participants were excluded if they had recently started to play music, or if they had more extensive musical training but no recent musical involvement. Listeners received nominal payment for participating in the experiment, which took approximately half an hour.

4.2. Apparatus

Stimuli were programmed on an IBM-XT personal computer using an adapted version of the *Adagio* software program developed by Roger Dannenberg at Carnegie–Mellon University. The computer was connected through a Roland MPU interface to a Yamaha TX816 FM (frequency modulation) tone generator. Stimuli were presented with a Yamaha amplifier (P2150) and a Yamaha 1204 MC series mixing console through a single JBL Model 4312A loudspeaker in a sound-attenuated room. Participants recorded their responses on the computer keyboard.

4.3. Stimulus materials

The eight melodic fragments from the British folk song collections are shown in Fig. 3. Four were in a major key and four in a minor key. Each fragment ended with an implicative interval meeting all of the following criteria: (1) the second tone of the implicative interval had equivalent or shorter duration relative to the first tone, (2) the second tone had a lower value in the tonal hierarchy of the key of the fragment than the first tone, (3) the second tone occurred on a metrically weaker beat than the first tone, (4) the second tone did not occur in the last or second-to-last position of a phrase, and (5) the second tone was 16–21 tones from the beginning of a phrase. The first four criteria ensured that the last two tones of each fragment were unclosed and truly implicative. The fifth criterion ensured that all fragments would be approximately equal in overall duration.

The stimuli were presented with a synthesized piano timbre at a tempo considered natural by the experimenter, and at a comfortable listening level adjusted according to each listener's preference. A subtle accent pattern

(i.e., small increases in intensity for strong beats) based on the time signatures of the fragments was used to clarify the metrical structure. One stimulus was altered slightly because none of the folk songs contained a downward implicative interval of 10 semitones that met all of the above criteria. Specifically, a downward interval of 12 semitones was identified, and the second tone was raised by 2 semitones to form a downward interval of 10 semitones. The altered song sounded stylistically consistent with the unaltered songs (experimenter's judgment).

4.4. Procedure

Listeners, who were tested individually, received instructions verbally and on the computer screen. They were told that they would hear fragments of melodies that began at the beginning of a phrase but ended in the middle of a phrase. Their task was to rate how well additional single tones (i.e., the test tones) continued the melodic fragments on a scale from 1 (extremely bad continuation) to 7 (extremely good continuation). They were told that each melodic fragment would still sound incomplete even with the test tone added; thus, they should not rate how well a test tone completed the fragments, but rather how well it continued the fragments. Finally, they were urged to use all seven points on the rating scale, limiting ratings of 1 and 7 to extreme cases.

Listeners began the experiment with a practice session, during which they heard a melodic fragment from the same folk song collections (but not a test fragment). The fragment was first presented three times; the fourth time a test tone was added to the end of the fragment. The duration of the test tone was equal to the duration of the tone at the same temporal location in the original song. Listeners made one rating for each of eight trials (i.e., eight different test tones) during the practice session. All trials (practice and test) were self-paced. After a listener had rated all eight test tones, the experimenter explained that the eight test tones represented a good cross-section of test tones in the actual experiment, and repeated the advice about the entire 7-point scale. The experimenter then left the room before the start of the test session.

The test session consisted of eight groups of trials, one for each of the melodic fragments. Each group was identical to the practice session except that listeners rated 15 rather than 8 different test tones. The 15 test tones consisted of all diatonic tones (i.e., all tones in the key of each fragment) within an octave up or down from the last tone of the fragment. For example, the test tones used for the first melodic fragment in Fig. 3 (ending on G_2 in the key of D major) were: G_2 , $F\#_2$, E_2 , D_2 , $C\#_2$, B_1 , A_1 , G_1 , $F\#_1$, E_1 , D_1 , $C\#_1$, B_0 , A_0 , and G_0 . The test tones were presented in a different random order for each listener and for each fragment; the order of the eight fragments was also randomized separately for each listener. Because listeners were required to make 120 ratings (15 test tones for each

of the eight fragments), they were allowed to take short breaks between fragments, as necessary.

5. Results

5.1. Agreement among listeners and listener-groups

To evaluate consistency across individuals, the data from each listener were correlated with those from every other listener. All 190 of the pairwise intersubject correlations were statistically significant ($N = 120$, $ps < .001$; mean $r = .605$). In these and in all subsequent tests, alpha levels were adjusted with the multistage Bonferroni correction for multiple tests (Darlington, 1990, pp. 249–276). For moderately trained listeners, the mean intersubject correlation was $.593$, $N = 120$, $p < .0001$; for listeners with limited training, it was $.640$, $N = 120$, $p < .0001$.

The consistency across listeners warranted the averaging of data for the primary analyses. For each of the 120 test tones, an average score for all 20 listeners was calculated (see Appendix A) as well as separate averages for listeners who were moderately trained and listeners with limited training. The simple correlation between the averaged ratings of listeners with moderate and limited levels of musical training was high, $r = .912$, $N = 120$, $p < .0001$, revealing very similar patterns of responding across the two groups.

5.2. I-R model

A preliminary analysis examined the independence of the model's predictor variables using the values corresponding to the 120 test tones the listeners rated. Correlations between each pairwise combination of predictors are provided in Table 1. After correcting for 15 tests, INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE were found to be significantly inter-correlated.

The next analysis examined how well the quantitative model (*average melodic continuation rating = weighted linear sum of quantified predictor variables plus a constant*), consisting of two dummy variables (REGISTRAL DIRECTION and INTERVALLIC DIFFERENCE) and four numerical variables (REGISTRAL RETURN, PROXIMITY, CLOSURE, and TONALITY), predicted listeners' average ratings. The model assumes that predictors are additive (no interactions). The results from the multiple regression analyses for listeners with limited training, listeners with moderate training, and all 20 listeners are presented in Table 2 (upper portion). The fit of the model to the data was highly significant for all three groups of listeners. After correcting for six tests, all five of the I-R model's predictor variables made significant unique contributions to the fit of the model across groups. The TONALITY

Table 1
Correlations between predictor variables in Experiment 1 ($N = 120$)

I-R model					
	Intervallic difference	Registral return	Proximity	Closure	Tonality
Registral direction	.101	.041	.009	.047	-.027
Intervallic difference		.000	.699**	.401**	.054
Registral return			.007	.117	.208
Proximity				.413**	.038
Closure					.016

Revised model			
	Registral return (revised)	Proximity (revised)	Tonality
Registral direction (revised)	.344*	.081	-.037
Registral return (revised)		.007	.015
Proximity (revised)			.041

* $p < .001$; ** $p < .0001$.

covariate was significant for listeners overall and for the group with limited training, but not for the moderately trained group.

As shown in Table 2, the I-R model explained approximately 65% of the variance in responses for each of the listener groups. However, the sum of the *unique* contributions of individual predictors to the fit of the model (i.e., the sum of the squared semipartial correlations) accounted for only about 25% of the variance in each analysis. Thus, approximately 40% of the variance in the data was explained redundantly (i.e., by more than one predictor variable).

The model was also fit to the data from each listener to test its consistency across individuals. The results revealed a significant fit to the data from each individual listener (corrected for 20 tests); highest $R^2 = .580$, lowest $R^2 = .252$, mean $R^2 = .451$, $N = 120$, $ps < .0001$. For each predictor variable, a pooled t -test compared the mean coefficient value for moderately trained listeners with that from listeners with limited training (corrected for six tests); no differences were observed.

5.3. Revised model

A revised model, containing three core principles rather than five, was derived from the data (i.e., by repeatedly attempting to explain the variance

Table 2

Multiple regression results for averaged ratings from Experiment 1 (sr^2 is the squared semipartial correlation; it represents the unique proportion of variance explained by a predictor variable)

	All 20 listeners	Listeners with moderate training (10)	Listeners with limited training (10)
<i>I-R Model</i>	$R^2 = .683$ $N = 120$ $p < .0001$	$R^2 = .634$ $N = 120$ $p < .0001$	$R^2 = .675$ $N = 120$ $p < .0001$
Registrational direction	$sr^2 = .029$ $p < .01$	$sr^2 = .035$ $p < .01$	$sr^2 = .020$ $p < .05$
Intervallic difference	$sr^2 = .024$ $p < .01$	$sr^2 = .020$ $p < .05$	$sr^2 = .024$ $p < .05$
Registrational return	$sr^2 = .067$ $p < .0001$	$sr^2 = .060$ $p < .0005$	$sr^2 = .068$ $p < .0001$
Proximity	$sr^2 = .089$ $p < .0001$	$sr^2 = .088$ $p < .0001$	$sr^2 = .082$ $p < .0001$
Closure	$sr^2 = .036$ $p < .005$	$sr^2 = .029$ $p < .05$	$sr^2 = .040$ $p < .005$
Tonality	$sr^2 = .015$ $p < .05$	$sr^2 = .009$ $p < .1$	$sr^2 = .019$ $p < .05$
<i>Revised model</i>	$R^2 = .759$ $p < .0001$	$R^2 = .725$ $p < .0001$	$R^2 = .728$ $p < .0001$
Registrational direction (revised)	$sr^2 = .084$ $p < .0001$	$sr^2 = .087$ $p < .0001$	$sr^2 = .074$ $p < .0001$
Proximity (revised)	$sr^2 = .472$ $p < .0001$	$sr^2 = .447$ $p < .0001$	$sr^2 = .456$ $p < .0001$
Registrational return (revised)	$sr^2 = .052$ $p < .0001$	$sr^2 = .050$ $p < .0001$	$sr^2 = .049$ $p < .0001$
Tonality	$sr^2 = .034$ $p < .0005$	$sr^2 = .025$ $p < .005$	$sr^2 = .041$ $p < .0001$

as parsimoniously as possible). These three principles were modifications of those described by the I-R model. REGISTRATIONAL DIRECTION was revised so that it applied only to large intervals, with large intervals implying a change in direction. As shown in Fig. 4, REGISTRATIONAL DIRECTION (REVISED) was coded 0 for small implicative intervals, 1 when a large implicative interval was followed by a realized interval that changed direction, and -1 when a large implicative interval was followed by a realized interval that continued in the same direction. REGISTRATIONAL RETURN was re-coded as an all-or-none variable (REGISTRATIONAL RETURN (REVISED)) such that instances where the second tone of the realized interval fell within two semitones of the first tone of the implicative interval were coded as 1 and all other instances were coded as 0.

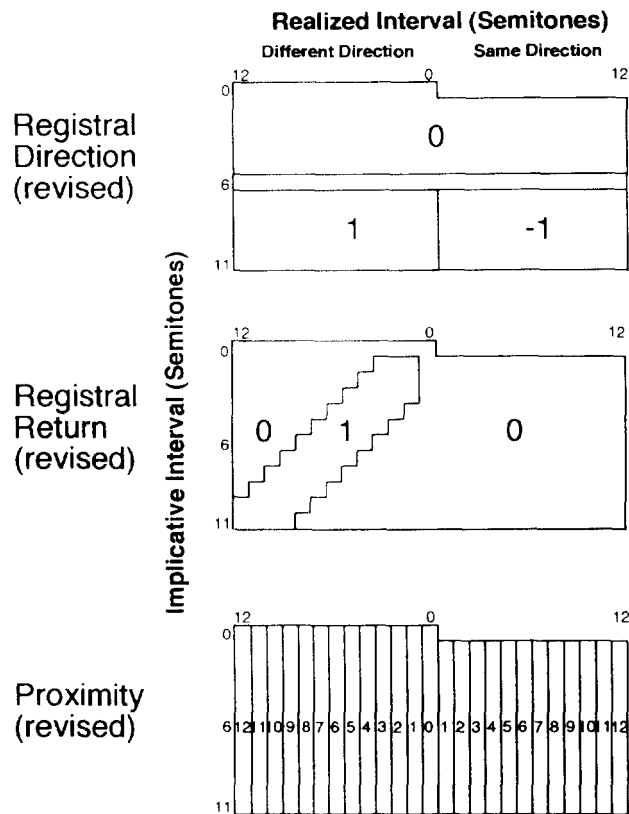


Fig. 4. The figures show how the three principles of the revised model were quantified in order to test the model.

The collinearity between INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE was handled by dispensing with INTERVALLIC DIFFERENCE and CLOSURE and by modifying PROXIMITY. PROXIMITY was revised to specify the size of the realized interval, such that PROXIMITY (REVISED) was coded 0 for realized intervals that were 0 semitones, 1 for realized intervals that were 1 semitone, 2 for 2 semitones, and so on. Because of this recoding, the model coefficient was expected to be negative rather than positive. As with the original model, the TONALITY covariate was included to control for effects of tonality.

A preliminary analysis examined the independence of the revised model's predictor variables. Pairwise correlations between predictors are provided in Table 1 (lower portion). After correcting for six tests, the revised model had one significant pairwise correlation, that between REGISTRAL DIRECTION (REVISED) and REGISTRAL RETURN (REVISED).

Results from multiple regression analyses using the revised model are provided in Table 2 (lower portion). For each listener group, the revised model predicted the data as well as the original model, with all four

predictors making significant unique contributions to the fit of the model (corrected for four tests for each group). Indeed, R^2 values were somewhat higher for the revised model for each of the three groups. Although there is no standard statistical test for comparing one model to another when one model is not a subset of the other, the ratio of the residual sum-of-squares from the original model to the residual sum-of-squares from the revised model has an F distribution (R.B. Darlington, personal communication, June, 1990). Using this statistical approach, the two models were compared separately for each listener group. The models did not differ significantly in their ability to predict the variation in the averaged data for any of the groups.

The revised model explained approximately 75% of the variance in the averaged responses for each listener group (see Table 2, lower portion). The unique contributions made by the four predictors accounted for approximately 62% of the variance in each case. Thus, with the revised model, only about 13% of the variance was explained redundantly.

As with the original model, the revised model was used to predict the data from each individual listener. The revised model produced a significant fit to the data from all listeners (corrected for 20 tests): highest $R^2 = .600$, lowest $R^2 = .251$, mean $R^2 = .497$, $N = 120$, $ps < .0001$. For each of the four predictor variables, a pooled t -test comparing the mean coefficient from listeners with limited training with that from listeners with moderate amounts of musical training (corrected for four tests) failed to reveal a difference between groups.

The final analysis compared individual listeners' R^2 s obtained with the original model to those obtained with the revised model. A Wilcoxon matched-pairs signed-ranks test (normal approximation) revealed that R^2 s were significantly higher for the revised model than they were for the original model, $z = 3.62$, $p < .0005$.

5.4. A subsidiary analysis

Six additional listeners with *no* formal music lessons (outside of school) were tested with the present set of stimulus materials (different laboratory and apparatus). The purpose was twofold: (1) to clarify the role of musical training, and (2) to confirm the validity of the revised model with a new data set. Patterns of responding were virtually identical to those found with the initial 20 listeners. Ratings from each listener were significantly correlated with averaged ratings from the 20 listeners in the main analysis (corrected for six tests, $ps < .0001$, $N = 120$, mean $r = .676$). The I-R model significantly predicted the data from each of the six untrained listeners ($N = 120$, $ps < .0001$, corrected for 6 tests, mean $R^2 = .358$), as did the revised model ($N = 120$, $ps < .0001$, corrected for 6 tests, mean $R^2 = .409$). Compared to the original model, the revised model produced higher R^2 values

across individual listeners, $z = 2.20$, $p < .05$ (Wilcoxon matched-pairs signed-ranks test, normal approximation).

6. Discussion

Although ratings in the context of melodic fragments from British folk melodies provided support for the I-R model, such ratings also prompted revisions in the model to decrease redundancy and the degree of collinearity among predictor variables. Judgments of how well different tones continued the melodic fragments were successfully predicted by the quantified aspects of the I-R model, but these judgments were predicted even more successfully by a simplified version of the model.

The primary analysis of the I-R model used multiple regression to predict the rating data as an additive weighted combination of the principles plus one covariate based on the tonal implications of the stimulus materials. The results supported the claims of the model in the following ways: (1) the model successfully predicted averaged ratings from all groups of listeners; (2) each predictor made a significant unique contribution to the fit of the model to the averaged ratings; (3) the model successfully predicted ratings from all individual listeners, regardless of the extent of their musical training; and (4) no differences due to extent of musical training were found, a finding consistent with Narmour's (1990, 1992) claim that melodic expectancies are based on general psychological principles that are not limited to listeners with extensive training in music.

The present findings also revealed the I-R model to be over-specified in its original form. Most of the variance in the data was explained in a redundant manner (i.e., by more than one predictor variable) because of the significant intercorrelations among three of the five predictor variables. Some collinearity among predictors is common in multiple regression models (Darlington, 1990, p. 150), however, and is not necessarily problematic (pp. 130–131). In fact, the inclusion of all of the predictors could be justified on the basis of each making a significant unique contribution to the fit of the model to the averaged data sets. Nonetheless, the redundancy and collinearity of the I-R model in its original form raise the possibility of simplification without sacrificing predictive power. Simplicity is one of the main criteria by which psychological models are evaluated (Cutting, Bruno, Brady, & Moore, 1992).

Indeed, a revised, simplified version of the model with fewer predictor variables than the original model and less collinearity maintained its predictive power. For each group of listeners, the revised model predicted the averaged data as well as the original model. Because of its simplicity and at least equal predictive power, the revised model could be considered superior to the original in accounting for the data reported in the present experiment (see Cutting et al., 1992). In fact, R^2 values from individual

listeners were higher for the revised model than for the original. The reduced collinearity among predictor variables also meant that each predictor of the revised model made relatively greater unique contributions to the model than did its counterpart (or counterparts) in the original model. Finally, the revised model and its predictors were consistent across differences in musical training, raising the possibility that the predictors of the revised model reflect general predispositions governing melodic expectancy, consistent with claims for the original predictors.

Because the stimulus fragments used in the present experiment were unambiguously tonal, it may seem surprising that the tonality covariate did not uniquely account for much of the variance in the data (less than 5% in all analyses of group data). Recall, however, that only tones in the key of the fragments were used as test tones to preclude the possibility that listeners might make their ratings primarily on the basis of key membership. Clearly, the tonality covariate would have accounted for a much larger unique portion of the variance if out-of-key test tones had also been rated.

Although listeners were requested to rate how well the test tones *continued* the melodic fragments rather than how well the tones *completed* the fragments, the present results cannot confirm that listeners were actually rating melodic continuations. Perhaps melodic continuations that sounded “bad” may have also sounded incomplete because additional tones were required to bring about a sense of stability and finality. In this sense, continuation and stability may have been confounded. Hence, the ratings may represent general “expectancy”, reflecting how well particular test tones continued and completed various melodic fragments.

Because the revised model was derived from the data reported here, it is not surprising that it fit the data well. A subsidiary analysis provided initial support for the validity of the revised model and its ability to predict other sets of data. If the revised model continues to outperform the original in subsequent tests, such evidence would confirm its psychological validity as well as its superiority as a model of melodic expectancy.

EXPERIMENT 2: JUDGMENTS OF ATONAL MELODIES

The present experiment was designed to test the applicability of the I-R model (Narmour, 1990, 1992) to atonal music. The development of an atonal style of composition (e.g., Berg, Schoenberg, Webern) began around the beginning of the 20th century as a reaction to the compositional constraints of 18th- and 19th-century tonal music. Unlike tonal music, atonal music uses no diatonic subset of tones from the chromatic scale, and no tone functions as a tonic from which listeners can evaluate the relative stability of other tones. One style of atonal music, 12-tone serial music, guarantees an equal distribution of the 12 chromatic tones through the use of a *tone row* (a particular ordering of the 12 chromatic tones, which must be completed

before any tone is repeated). Although atonal music has stimulated a large literature in music theory and analysis, it remains unfamiliar to most Western listeners because of the domination of tonal music from the 18th and 19th centuries in concert programs, broadcasting, and music education. Folk and popular music have also retained the structures of Western tonal music. Thus most listeners, because of their limited exposure to atonal music, are unlikely to have implicit knowledge of its structure.

The stimuli in the present experiment were extracts from *Lieder* composed by Anton Webern (1921, 1923, 1924), which were presented in a task identical to that of Experiment 1. The vocal melodies, shown in Fig. 5, were selected from Webern's pre-serial period (Opus 3, Opus 4, and Opus 15) so that they would be atonal but not in the serial style. (With serial music, constraints of the tone row may have interfered with melodic expectancies.) The vocal melodies were originally written with simple accompaniments: solo piano in five instances, flute and clarinet in one instance, clarinet,

Fragment 1

Fragment 2

Fragment 3

Fragment 4

Fragment 5

Fragment 6

Fragment 7

Fragment 8

Fig. 5. The figure shows the melodic fragments used in Experiment 2. The fragments are taken from Webern *Lieder*.

trumpet, and viola in one instance, and solo flute, bass clarinet, trumpet, harp, and viola in one instance. For consistency with Experiment 1, the accompaniments were not included in the present experiment. Again, small and large implicative intervals (in both ascending and descending forms) were chosen to be representative of Narmour's (1990, 1992) classifications. The small implicative intervals in the present experiment were 1 semitone and 4 semitones; the large implicative intervals were 8 semitones and 11 semitones. Test tones consisted of all chromatic scale tones within an octave of the last tone of each fragment.

Because the stimulus materials were atonal, no predictor variable was included to control for effects of tonality. Thus, the present experiment provided an opportunity to test the principles of the implication-realization model in stimulus contexts where learned, style-specific influences of tonality would be negligible.

7. Method

7.1. Participants

The participants were 26 members of the Cornell University community; 13 were *musically trained*, 13 others were *untrained*. The musically trained listeners were recruited from an upper-level course in the psychology of music and had extensive training in music (mean of 11.8 years of music lessons). The untrained listeners had a mean of only 2.69 years of music lessons and had not taken music lessons or regularly played music within the previous 2 years. Listeners received course credit or nominal payment for participating in the experiment, which took approximately 1 hour.

7.2. Apparatus

The apparatus was identical to that of Experiment 1.

7.3. Stimulus materials

Eight melodic fragments were selected from the Webern *Lieder* (see Fig. 5). These fragments ended with implicative intervals that met four of the five criteria outlined in Experiment 1; the criterion relating to tonal stability was obviously inapplicable to atonal fragments. The stimuli were presented with a synthesized piano timbre. Tempo was based on metronome markings on the musical scores and what was natural-sounding to the experimenter. The fragments were presented at a comfortable listening level, adjusted according to listeners' preferences. Subtle differences in intensity were used to help clarify the metrical structure of the melodic fragments, as indicated by the time signatures.

7.4. Procedure

The procedure was identical to Experiment 1 except that, during the testing session, listeners rated 25 different test tones for each melodic fragment. The 25 test tones represented all chromatic tones within an octave of the second tone of the implicative interval. For example, the 25 test tones used for Fragment 1 in Fig. 5 (ending on F#₄) were: F#₅, F₅, E₅, D#₅, D₅, C#₅, C₅, B₄, A#₄, A₄, G#₄, G₄, F#₄, F₄, E₄, D#₄, D₄, C#₄, C₄, B₃, A#₃, A₃, G#₃, G₃, and F#₃. The 25 test tones were presented in a different random order for each listener for each fragment, and the eight fragments were presented in a different random order for each listener. Each listener was required to provide 200 ratings during the test session (25 test tones for each of 8 fragments).

8. Results

8.1. Agreement among listeners and listener-groups

The data from each listener were correlated with those from every other listener (corrected for 325 tests): 257 of the 325 pairwise intersubject correlations were significant at a corrected .05 alpha-level. The mean intersubject correlation was .371, $N = 200$, $p < .0001$. Among trained listeners, the mean intersubject correlation was .324, $N = 200$, $p < .0001$, and 52 of the 78 pairwise intersubject correlations were significant (corrected for 78 tests). Among untrained listeners, the mean intersubject correlation was .442, $N = 200$, $p < .0001$, and all of the 78 intersubject correlations were significant ($ps < .005$, corrected for 78 tests).

Further examination of the intersubject correlations revealed that all 26 of the nonsignificant correlations for trained listeners involved four specific listeners. Hierarchical clustering and multidimensional scaling solutions of intersubject correlations showed these four listeners to be outliers among the trained listeners, and largely dissimilar from one another. Because significant intersubject agreement was limited to untrained listeners and to 9 of the 13 trained listeners, the data were averaged across these listeners (excluding the four outliers) for further analyses. Thus, for each of the 200 test tones that were rated by listeners, an average score for 22 listeners was obtained (see Appendix B) as well as an average score for 9 trained listeners and an average score for 13 untrained listeners. The simple correlation between the averaged ratings of trained and untrained listeners was .847, $N = 200$, $p < .0001$.

8.2. I-R model

The model was coded as in Experiment 1 except for tonality. The intercorrelations between predictor variables (as coded for the 200 test

Table 3
Correlations between predictor variables in Experiment 2 ($N = 200$)

I-R model				
	Intervallic difference	Registral return	Proximity	Closure
Registral direction	.120	.026	-.000	.028
Intervallic difference		.007	.639*	.375*
Registral return			-.014	.117
Proximity				.370*

Revised model		
	Registral return (revised)	Proximity (revised)
Registral direction (revised)	.326*	-.049
Registral return (revised)		.014

* $p < .0001$.

tones) are provided in Table 3 (upper portion). The results were identical to those found in Experiment 1, INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE being significantly intercorrelated (after correcting for 10 tests).

The multiple regression results are provided in Table 4 (upper portion). The fit of the model to the data was highly significant for musically trained listeners (excluding the four outliers), untrained listeners, and the two groups combined. The predictor variables showed fairly consistent effects in the three analyses. Three predictors (REGISTRAL DIRECTION, REGISTRAL RETURN, and PROXIMITY) made significant unique contributions to the model for all three listener groups. INTERVALLIC DIFFERENCE was significant for musically trained listeners but not for untrained listeners or for listeners overall. CLOSURE was significant for untrained listeners but not for the other listener groups.

As shown in Table 4 (upper portion), the model explained approximately 45% of the variance in the data for each analysis. However, the unique contributions of the individual predictor variables (calculated by summing the squared semipartial correlations) accounted for only about half of the explained variance in each case. Thus, between 20% and 25% of the variance was consistently being explained redundantly, indicating over-specification in the model.

The model was also fit to the data for each listener to evaluate its consistency across individuals (corrected for 26 tests). The model produced a significant fit to the ratings for all listeners except for the four designated previously as outliers ($ps < .005$). For the 22 listeners whose responses were

Table 4

Multiple regression results for averaged ratings from Experiment 2 (sr^2 is the squared semipartial correlation: it represents the unique proportion of variance explained by a predictor variable)

	All 22 listeners	Musically trained listeners (9)	Musically untrained listeners (13)
<i>I-R Model</i>	$R^2 = .461$ $N = 200$ $p < .0001$	$R^2 = .413$ $N = 200$ $p < .0001$	$R^2 = .459$ $N = 200$ $p < .0001$
Registral direction	$sr^2 = .093$ $p < .0001$	$sr^2 = .056$ $p < .0005$	$sr^2 = .116$ $p < .0001$
Intervallie difference	$sr^2 = .017$ $p < .05$	$sr^2 = .028$ $p < .01$	$sr^2 = .009$ $p < .1$
Registral return	$sr^2 = .035$ $p < .005$	$sr^2 = .022$ $p < .05$	$sr^2 = .042$ $p < .0005$
Proximity	$sr^2 = .071$ $p < .0001$	$sr^2 = .070$ $p < .0001$	$sr^2 = .065$ $p < .0001$
Closure	$sr^2 = .016$ $p < .05$	$sr^2 = .006$ p n.s.	$sr^2 = .023$ $p < .01$
<i>Revised model</i>	$R^2 = .527$ $p < .0001$	$R^2 = .487$ $p < .0001$	$R^2 = .509$ $p < .0001$
Registral direction (revised)	$sr^2 = .094$ $p < .0001$	$sr^2 = .056$ $p < .0001$	$sr^2 = .116$ $p < .0001$
Proximity (revised)	$sr^2 = .358$ $p < .0001$	$sr^2 = .379$ $p < .0001$	$sr^2 = .307$ $p < .0001$
Registral return (revised)	$sr^2 = .016$ $p < .05$	$sr^2 = .011$ $p < .05$	$sr^2 = .018$ $p < .01$

successfully predicted, the highest R^2 was .433, the lowest R^2 was .117, and the mean R^2 was .252 ($N = 200$; $p < .0001$, $p < .005$, and $p < .0001$, respectively). For each of the six predictors, pooled t -tests (corrected for six tests) were used to compare mean coefficient values for musically trained and untrained listeners (including the outliers); none of the differences was significant.

8.3. Revised model

The revised model derived from the data in Experiment 1 was used to predict the present data set. A preliminary analysis examined the intercorrelations between pairwise combinations of predictor variables (see Table 3, bottom portion), corrected for three tests. As before, REGISTRAL DIRECTION (REVISED) and PROXIMITY (REVISED) were significantly correlated. The other two pairwise correlations were not significant.

The revised model was used to predict the averaged data from the three

listener groups. The results are provided in Table 4 (bottom portion). For each group, the revised model showed a highly significant fit to the data. Actual R^2 values from the revised model were somewhat higher than the corresponding values from the original model, although the models did not differ significantly in their predictive power for any of the listener groups. For each group, all three revised principles made a significant unique contribution to the fit of the model.

The revised model explained approximately 50% of the variation in the averaged data for each group of listeners (see Table 4, lower portion). The unique contributions of each predictor variable explained approximately 45% of the variance in each case, with only about 5% of the data explained redundantly.

The revised model was also used to predict the data from each of the 26 listeners (corrected for 26 tests). As with the original model, the revised model significantly predicted responses from all listeners except for the four outliers ($ps < .005$). For the 22 listeners for which the model was successful, the highest R^2 was .431, the lowest R^2 was .092, and the mean R^2 was .272 ($N = 200$, $p < .0001$, $p < .005$, and $p < .0001$, respectively). For each of the three predictor variables, an independent samples *t*-test comparing mean coefficients from musically trained and untrained listeners (including the outliers) found no difference between groups (corrected for three tests).

The final analysis compared individual listeners' R^2 values from the original model with those from the revised model (outliers included). A Wilcoxon matched-pairs signed-ranks test (normal approximation) revealed that the revised model provided a better explanation of the variation in individual listeners' responses than did the original model, $z = 2.03$, $p < .05$.

8.4. Outliers

Test-tone ratings from the four outliers were analyzed as a function of the total duration of each test tone in the preceding melodic fragment and as a function of how recently the test tone occurred in the fragment. Krumhansl, Sandell, and Sergeant (1987) reported that these variables affected their listeners' ratings of test tones in atonal contexts. After correcting for four tests, tone duration was not significantly correlated with ratings from any of the outliers. By contrast, tone recency was significantly correlated with ratings from all four outliers (corrected for four tests). Two of the outliers tended to give higher ratings to test tones that appeared more recently in the melodic fragments, $r = .183$, $N = 200$, $p < .05$, and $r = .242$, $N = 200$, $p < .005$; the other two outliers tended to give lower ratings to test tones that appeared more recently, $r = -.150$, $N = 200$, $p < .05$, and $r = -.277$, $N = 200$, $p < .0005$. The proportion of variation in the outlier's data explained by tone recency was much smaller (less than 5% on average) than the average amount of variation explained in the other listeners' data by either the I-R model or the revised model. Neither tone duration nor tone recency

significantly improved the fit of the original or revised models to the averaged ratings from the other 22 listeners.

9. Discussion

The results of the present experiment were similar to those found in Experiment 1. In general, judgments of melodic continuations were significantly predicted by the I-R model for both musically trained and untrained listeners. Moreover, the revised model, which had been derived from the data of Experiment 1, provided a better and simpler explanation of the variation in listeners' responses in the present experiment than did the original model. Thus, these results support the claim that the I-R model can be revised and simplified without loss of predictive power.

Nevertheless, the predictive power of the original and revised models was weaker in the present experiment than in Experiment 1. Four of the musically trained listeners produced judgments that were dissimilar from other listeners and from each other; in fact, their ratings were not significantly predicted by either model. With the exception of these outliers, however, intersubject agreement was consistent but considerably weaker than in Experiment 1. Presumably, this finding was a consequence of the atonal stimulus materials, as was the poorer overall fit of the model relative to Experiment 1 (i.e., lower multiple R^2 's). The lack of a tonic tone, or reference point, in these melodies would have increased the difficulty of processing such materials (Cuddy, Cohen, & Mewhort, 1981; Cuddy, Cohen, & Miller, 1979). From Garner's (1970, 1974) perspective, the atonal fragments could be considered "poor" auditory patterns, evoking many possible alternatives because of their non-adherence to diatonic scale structure (Bartlett & Dowling, 1988). Thus, the sets of implied continuations for these fragments may have been larger than the corresponding sets in Experiment 1, resulting in increased variation (more noise) in response patterns.

The differences in processing difficulty between tonal and atonal stimulus materials are not accounted for by the I-R model, yet a comparison of the results of the present experiment to those of Experiment 1 suggests that the core principles of the I-R and revised models exert a stronger influence in tonal rather than atonal contexts. Thus, expectancies governed by fundamental melodic principles may be weaker in contexts that are, in general, relatively incoherent. The influence of the core principles may also have been affected by other specific characteristics of the stimulus contexts. For example, proximity may have been relatively weak in the present experiment because successive tones in the atonal melodies were, on average, less proximate than those in Experiment 1 (see Figs. 3 and 5). Another possibility is that the overall increase in unexplained variation stemmed, in part, from the fewer number of predictor variables in the model (three as

opposed to four), or from the greater number of test tones included in the present experiment (all chromatic scale tones as opposed to those of a diatonic subset) and, hence, more degrees of freedom in the analyses.

For all groups of listeners (musically trained excluding the four outliers, musically untrained, and combined), however, both the I-R model and the revised model were highly significant, and both models significantly predicted ratings from each individual listener (excluding the four outliers). Thus, the quantified predictors of both models significantly accounted for the variation in melodic continuation judgments, even in atonal contexts, both at the level of groups of listeners and at the level of individual listeners. This result implies that listeners can transfer their processing strategies for familiar melodies to unfamiliar musical styles.

As in Experiment 1, the collinearity of the principles of the I-R model generated considerable redundancy in terms of the unique contributions of the individual predictors to the model. Specifically, *INTERVALLIC DIFFERENCE*, *PROXIMITY*, and *CLOSURE* were highly intercorrelated, making their unique contributions to the model relatively weak. The revised model replaced these three with a sole predictor variable, *PROXIMITY (REVISED)*, which was not correlated with the revised model's other two predictors. The other two (*REGISTRAL DIRECTION (REVISED)* and *REGISTRAL RETURN (REVISED)*) were significantly correlated, however, as they were in Experiment 1, raising the possibility that the revised model might be simplified further in the future.

EXPERIMENT 3: JUDGMENTS OF NON-WESTERN TONAL MELODIES

The present experiment was designed to test the applicability of the I-R model (Narmour, 1990, 1992) to a non-Western musical style, namely Chinese folk melodies. Much of Chinese folk music is pentatonic, that is, composed from a scale consisting of five tones. In contrast, most Western tonal music is composed from seven-tone major and minor scales. The Chinese pentatonic scale contains the relative pitch relations among tones represented by the black keys on the piano; it is formed by starting on any tone and including tones successively 2, 3, 2, 2, and 3 semitones higher. The scale, which has five modes beginning on each of the five tones of the scale (Koon, 1979; Laloy, 1979), is similar in some respects to Western scales. Intervals between any two tones are integer multiples of semitones, as is the case with Western scales. The move towards equal temperament (i.e., an octave that is divided into 12 equally distant semitones) began in China as early as the fourth century B.C. (Yung, 1980).

There are fundamental differences, however, between the Chinese pentatonic scale and Western major and minor scales. Because two adjacent tones in the scale may be three semitones apart, this interval constitutes a single step in Chinese melody but a leap in Western melody (Koon, 1979).

Because of the structure of the pentatonic scale, no two tones in the scale form intervals of 1, 6, or 11 semitones, all of which are found in Western major and minor scales. Another difference is that the Chinese pentatonic scale does not exhibit the tonal structure of Western music, but rather a “quasi-tonality” determined by the distribution of tones in a musical piece (Gilman, 1892). A Chinese musical piece often ends with the tone that is most frequently sounded in the piece. This last tone usually corresponds to the tone on which the mode itself begins and might, therefore, be considered the tonic of the piece (Koon, 1979).

In Chinese music theory, three of the five tones in the pentatonic scale form a *basis set*, which consists of the tonic and two other tones (Sin-yan, 1979, 1981). Tonality is, therefore, determined by whether or not a given tone is a member of the basis set. In four of five modes, a second member of the basis set is usually the tone seven semitones above the tonic (Shu-hsien, 1986). In the one remaining mode (corresponding to the black keys of the piano beginning on A#), the tone seven semitones above the A# (i.e., E# or F) is absent from the scale, so the tone eight semitones above the A# (i.e., F#) is included in the basis set instead. The third member of the basis set is a tone that is either three, four, or five semitones above the tonic. The tones that form the basis set for a particular mode can vary from one region to another (Sin-yan, 1979). The basis set sometimes corresponds to Western major or minor triads. Often, however, the basis set contains the tonic and the tones five semitones and seven semitones above the tonic and does not correspond to any Western harmony. Chinese music also differs from Western music in that the interval of five semitones (perfect fourth) plays an important role in tonality, in addition to intervals of three and four semitones (minor and major thirds) (Shu-hsien, 1986; Sin-yan, 1981). In Western harmony, simultaneous tones are grouped primarily in intervals of three and four semitones.

Although the extent to which Western listeners are sensitive to these structural aspects of Chinese music is unknown, the implicit assumption is that extensive experience is required for the internalization of tonal structures. Numerous studies that reveal effects of development and musical training (e.g., Krumhansl & Keil, 1982; Morrongiello & Roes, 1990; Morrongiello, Roes, & Donnelly, 1989; Trainor & Trehub, 1992) generally support this view. Nonetheless, Chinese and Western musical systems have a number of common features, as noted. Moreover, Western listeners are sensitive to some tonal structures in non-Western music. In studies with North Indian (Castellano, Bharucha, & Krumhansl, 1984) and Indonesian (Kessler, Hansen, & Shepard, 1984) stimulus materials, Western listeners produced responses consistent with theoretical descriptions of the relevant non-Western style. Features of the stimuli, such as tone repetition, duration, and serial position may have acted as cues to their tonal organization. Thus, Western listeners in the present experiment may also show response patterns that reflect sensitivity to Chinese tonal structure.

The melodic materials for the present experiment (see Fig. 6) were selected from a collection of Chinese folk songs (*Chung-kuo min kuo hsuan* (Chinese Folk Songs), People's Music Publishing Company, 1980). The participants, who were recruited without regard to musical training, included native Chinese listeners as well as listeners born and raised in the United States. Thus, musical enculturation rather than training was an issue in the present experiment. The set of test tones consisted of all tones in the pentatonic scale of the mode of the fragment. As in Experiment 1, the exclusion of test tones from outside the scale precluded the possibility of rating strategies based primarily on whether or not a given test tone was a member of the scale. The reduced number of test tones per fragment (compared to Experiments 1 and 2) permitted an increase in the number of fragments. The fragments selected ended with small intervals (2, 3, and 4 semitones) or large intervals (8, 9, and 10 semitones) in ascending and descending versions for each interval, yielding a total of 12 fragments. For each fragment, the experimenter and an ethnomusicologist specializing in Eastern Asian music identified the tones comprising the basis set (M. Hatch, personal communication, May, 1991). The predictor variable TONALITY was coded as a dummy variable (i.e., 1 for test tones belonging to the basis set, 0 otherwise).

10. Method

10.1. Participants

The participants were 16 members of the Cornell University community, 8 of whom were born and raised in the People's Republic of China and 8 others born and raised in the United States. The Chinese listeners, who had resided in America for an average of 2 years, 7 months (ranging from 2 weeks to 4 years, 8 months), reported that they had grown up listening to Chinese music. One Chinese listener also reported exposure to Western music during her childhood. Listeners received course credit or token remuneration for their participation, which took approximately 40 minutes.


10.2. Apparatus

The apparatus was identical to that used in Experiment 1.

10.3. Stimulus materials

Twelve melodic fragments were selected from the pentatonic folk songs in the Chinese folk song collection (Fig. 6). The fragment-final implicative intervals met the same criteria as in Experiment 1 with the following exceptions: (1) the criterion of relative tonal stability of the two tones

Fragment 1



Fragment 2



Fragment 3



Fragment 4



Fragment 5



Fragment 6



Fragment 7



Fragment 8



Fragment 9



Fragment 10



Fragment 11



Fragment 12



Fig. 6. The figure shows the melodic fragments used in Experiment 3. The fragments are taken from Chinese folk songs

making up the implicative interval was dropped because the stimulus materials were from a non-Western system of tonality, and (2) the second tone of the implicative interval was 13–21 tones from the beginning of a phrase. The stimuli were presented with a synthesized piano timbre at a natural-sounding tempo (experimenter's judgment) and at a comfortable listening level (adjusted for each listener). Subtle differences in intensity were used to clarify the metrical structure of the melodic fragments.

10.4. Procedure

The procedure was identical to that of Experiment 1 except that listeners received 12 groups of trials, 1 for each melodic fragment, and rated 11 different test tones for each fragment. The 11 test tones represented all tones in the scale of the stimuli within an octave up or down from the last tone of the melodic fragment. For example, the 11 test tones used for the melodic fragment shown in Fig. 6, Fragment 1 (ending on G_4) were: G_5 , F_5 , D_5 , C_5 , A_4 , G_4 , F_4 , D_4 , C_4 , A_3 , and G_3 . The 11 test tones were presented in a different random order for each fragment and listener; the 12 fragments were also presented in a different random order for each listener. Each listener made 132 ratings during the test session (11 test tones for each of 12 fragments).

11. Results

11.1. Agreement among listeners and listener-groups

The data from each listener were correlated with those from every other listener (corrected for 120 tests). The mean intersubject correlation was .572, $N = 132$, $p < .0001$. All 120 of the pairwise intersubject correlations were statistically significant after correcting for multiple tests ($N = 132$, $ps < .0001$). The mean intersubject correlation among American listeners was .619, $N = 132$, $p < .0001$; among Chinese listeners it was .550, $N = 132$, $p < .0001$.

Consistent intersubject agreement warranted the averaging of data across listeners for the primary analyses. For each of the 132 test tones that were rated, an average score for all 16 listeners was obtained (see Appendix C), as well as separate average scores for American and Chinese listeners. The simple correlation between American and Chinese listeners' average ratings was .884, $N = 132$, $p < .0001$.

11.2. I-R model

The predictor variables of the I-R model were coded as in Experiment 1 except that TONALITY was coded as a dummy variable (reflecting membership

in the basis set). A preliminary analysis examined the intercorrelations between predictors of the model (corrected for 15 tests), using the values corresponding to the set of 132 tones rated in the present experiment. As in Experiments 1 and 2, the predictors INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE were significantly intercorrelated (see Table 5, upper portion).

Results from multiple regression analyses for the data averaged over Chinese listeners, American listeners, and all 16 listeners are presented in Table 6 (upper portion). The fit of the model to the data was highly significant for all three listener groups ($ps < .0001$). After correcting for multiple tests, four predictor variables (INTERVALLIC DIFFERENCE, REGISTRAL RETURN, PROXIMITY, and CLOSURE) made significant unique contributions to the model for each group of listeners. REGISTRAL DIRECTION and TONALITY were not significant for any of the groups.

For each listener group, the I-R model accounted for about 65% of the variation in the data (see Table 6, upper portion). The sum of the unique contributions of individual predictor variables explained less than 25% of the total variance in each case. Thus, more than 40% of the variance was explained redundantly for each of the groups.

The multiple regression model produced a significant fit to the data from each individual listener (corrected for 16 tests), highest $R^2 = .618$, lowest $R^2 = .305$, mean $R^2 = .442$, $N = 132$, $ps < .0001$. To test for cross-cultural

Table 5
Correlations between predictor variables in Experiment 3 ($N = 132$)

I-R Model					
	Intervallic difference	Registral return	Proximity	Closure	Tonality
Registral direction	.107	.009	-.027	.042	.015
Intervallic difference		.047	.628*	.392*	-.004
Registral return			.003	.103	.086
Proximity				.503*	-.007
Closure					-.014
Revised model					
	Registral return (revised)	Proximity (revised)	Tonality		
Registral direction (revised)	.310*	-.093	.009		
Registral return (revised)		.013	.031		
Proximity (revised)			-.004		

* $p < .0001$

Table 6

Multiple regression results for averaged ratings from Experiment 3 (sr^2 is the squared semipartial correlation; it represents the unique proportion of variance explained by a predictor variable)

	All 16 listeners	Chinese listeners (8)	American listeners (8)
<i>I-R Model</i>	$R^2 = .690$ $N = 132$ $p < .0001$	$R^2 = .644$ $N = 132$ $p < .0001$	$R^2 = .662$ $N = 132$ $p < .0001$
Registral direction	$sr^2 = .010$ $p < .1$	$sr^2 = .012$ $p < .1$	$sr^2 = .007$ p n.s.
Intervallic difference	$sr^2 = .065$ $p < .0001$	$sr^2 = .040$ $p < .005$	$sr^2 = .085$ $p < .0001$
Registral return	$sr^2 = .031$ $p < .005$	$sr^2 = .030$ $p < .01$	$sr^2 = .029$ $p < .005$
Proximity	$sr^2 = .067$ $p < .0001$	$sr^2 = .087$ $p < .0001$	$sr^2 = .043$ $p < .0005$
Closure	$sr^2 = .048$ $p < .0001$	$sr^2 = .043$ $p < .001$	$sr^2 = .048$ $p < .0005$
Tonality	$sr^2 = .009$ $p < .1$	$sr^2 = .004$ p n.s.	$sr^2 = .013$ $p < .1$
<i>Revised model</i>	$R^2 = .755$ $p < .0001$	$R^2 = .701$ $p < .0001$	$R^2 = .723$ $p < .0001$
Registral direction (revised)	$sr^2 = .077$ $p < .0001$	$sr^2 = .074$ $p < .0001$	$sr^2 = .070$ $p < .0001$
Proximity (revised)	$sr^2 = .570$ $p < .0001$	$sr^2 = .539$ $p < .0001$	$sr^2 = .535$ $p < .0001$
Registral return (revised)	$sr^2 = .015$ $p < .05$	$sr^2 = .010$ $p < .1$	$sr^2 = .019$ $p < .01$
Tonality	$sr^2 = .010$ $p < .05$	$sr^2 = .005$ p n.s.	$sr^2 = .014$ $p < .01$

differences in the strength of each predictor variable, pooled *t*-tests were used to compare mean coefficient values for American listeners with those for Chinese listeners (corrected for six tests). No significant differences were found.

11.3. Revised model

The revised model, derived from the data of Experiment 1, was used to predict the data from the present experiment. A preliminary analysis examined the intercorrelations between pairs of predictor variables (corrected for six tests), which are provided in the lower portion of Table 5. As

in Experiments 1 and 2, a significant correlation was found between REGISTRAL DIRECTION (REVISED) and REGISTRAL RETURN (REVISED).

The revised model was used to predict averaged ratings from listeners overall as well as from Chinese and American listeners. Results from the multiple regression analyses are provided in Table 6 (lower portion). The model successfully predicted averaged ratings for each of the three listener groups. Although actual R^2 values were somewhat higher for the revised model than for the original, this difference was not statistically significant for any of the groups. All four predictor variables made significant unique contributions in explaining the variance for American listeners and for listeners overall. For Chinese listeners, the three core predictors were significant but the TONALITY covariate was not.

For each listener group, the revised model explained between 70% and 75% of the variation in averaged ratings. The sum of the unique contributions of individual predictor variables accounted for between 60% and 65% of the variation. Thus, for each group, only about 10% of the variation in the averaged data was explained redundantly with the revised model.

The revised model was also fit to individual listener's ratings, producing a significant fit for each listener (corrected for 16 tests); highest $R^2 = .600$, lowest $R^2 = .299$, mean $R^2 = .473$ ($ps < .0001$). Pooled t -tests, conducted separately for each predictor variable (corrected for four tests), compared mean coefficients between Chinese and American listeners. No differences were found. Finally, a Wilcoxon matched-pairs signed-ranks tests (normal approximation) revealed that the revised model explained more variance across individual listeners than did the original model, $z = 2.53$, $p < .05$.

12. Discussion

The results of the present experiment replicated those of Experiments 1 and 2. Because the stimuli used in the present experiment were taken from Chinese folk melodies, the findings of Experiments 1 and 2 were shown to extend to melodies from a non-Western musical culture. Although judgments of melodic continuations were significantly predicted by the I-R model across all listeners and listener groups, the revised version of the model provided a simpler explanation of response patterns with no loss of predictive power.

The multiple correlations resulting from both models were comparable to those of Experiment 1 (as were the number of test tones and the degrees of freedom in the analyses) and were highly significant for each of the listener groups (Chinese, American, and combined) and for each individual listener. Thus, the degree of inter-listener consistency was similar to that in Experiment 1, despite the cultural differences between listeners. Although

there may be systematic differences in responding due to cultural background, none was uncovered in the present analyses. Correlations between pairs of listeners showed, moreover, that the responses from any listener could explain about one-third of the variation in responses from any other listener, regardless of cultural background. Thus, although much of the variance in the data is due to individual differences, these differences were similar in magnitude whether or not listeners were raised in the same musical culture. The cross-cultural similarities revealed here indicate that the experimental method used in the present experiment (and in Experiments 1 and 2) may be ideal for tapping general principles of melodic expectancy, such as those suggested by Narmour (1990, 1992).

As in Experiments 1 and 2, collinearity among predictor variables of the I-R model resulted in relatively small unique contributions of INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE in explaining the variation in the data. The revised model reduced the amount of collinearity among predictors and consequent redundancy, providing for a simpler model without loss of predictive power. Although multiple R^2 s were higher for each listener group, the differences were not statistically significant. Nevertheless, multiple R^2 s were significantly higher across individual listeners.

With the revised model, TONALITY was a significant predictor of averaged ratings from American listeners but not of those from Chinese listeners. This finding suggests that a hierarchical differentiation of scale tones might be more characteristic of Western than of Chinese tonal schemas. Across individual listeners, however, the strength of the TONALITY predictor did not differ due to listeners' cultural background. With the original I-R model, moreover, TONALITY was nonsignificant for each listener group. Thus, the simplest interpretation of these findings is that the TONALITY covariate was a relatively weak predictor of listeners' ratings throughout the present experiment, as it was in Experiment 1. Recall that only tones belonging to the scale of each stimulus fragment were used as test tones. It is not surprising, then, that the influence of tonality, due strictly to within-scale differences in stability, was relatively weak.

The melodic continuation ratings of Experiments 1, 2, and 3 were remarkably consistent across the three experiments. Individual differences as a function of musical or cultural background were minimal, except for a few musically trained listeners who responded idiosyncratically to atonal melodies (Experiment 2). Nevertheless, definitive evidence of the universality of the underlying principles of the I-R model would require evaluation of the applicability of such principles across all musical styles and listeners. Despite the restricted sampling of musical styles and listeners (as well as the superior performance of the revised model compared to the original model), the results from Experiments 1, 2, and 3 indicate that the I-R model provides an excellent starting point in the search for general psychological principles governing expectancy in melody.

REANALYSIS OF CARLSEN (1981)/UNYK AND CARLSEN (1987) PRODUCTION DATA

The final test of the I-R model (Narmour, 1990, 1992) involved a reanalysis of data from other investigators. Carlsen (1981) and Unyk and Carlsen (1987) presented two-tone stimulus intervals as the beginning of a melody, and asked listeners to sing tones that might continue the melody. The stimulus intervals were multiples of semitones, ranging from a descending interval of 12 semitones to an ascending interval of 12 semitones, including unisons (a total of 25 stimulus intervals). Carlsen (1981) and Unyk and Carlsen (1987) restricted their analyses to the first tone of the sung responses (coded with respect to the second tone of the stimulus interval). They also excluded continuation tones that were farther than an octave from the second tone of the stimulus interval, but such instances were rare.

Carlsen (1981) tested students from professional music schools in Germany, Hungary, and the United States. These listeners were presented with stimulus intervals in their vocal range, and responded 15 times to each of the stimulus intervals. Musical training and voice register had no effect on the pattern of responses. There were effects of cultural background, however. German and Hungarian listeners differed on 5 out of 25 stimulus intervals, German and American listeners differed on 7 out of 25 stimulus intervals, and Hungarian and American listeners differed on 18 out of 25 stimulus intervals. Although these differences occurred more frequently than would be expected by chance, they could not be readily attributed to particular differences in musical style. Moreover, actual differences may have been masked by listeners' extensive training and familiarity with Western art music from the 18th and 19th centuries (the "common practice" period).

Unyk and Carlsen (1987) repeated the procedure with American musicians, who made only 5 responses to each of the 25 two-tone combinations. The goal of this follow-up study was to obtain individual expectancy profiles so that, for each stimulus, a strongly expected tone, a weakly expected tone, and an unexpected tone could be identified for each listener. Each listener's individual profile was then used in tasks that tested melodic recall, perception, and identification as a function of expectancy strength and expectancy fulfillment/denial. The results indicated that unexpected tones generated more errors than expected tones.

The data of Carlsen (1981) and Unyk and Carlsen (1987) can also function as a test of the I-R model, providing a source of convergent evidence for the findings in Experiments 1, 2, and 3. Although their two-tone stimulus intervals lacked the complexity and naturalness of the musical fragments in the present report, they are nevertheless implicative (unclosed) as defined by the I-R model (Narmour, 1990, 1992). Their intervals were unclosed durationally (each tone was the same duration), metrically (the first tone would tend to be perceived as on a stronger beat than the second tone), and harmonically (the first tone, perceived as the

tonic, would be more stable than the second tone). Whereas the two-tone stimuli can be considered implicative intervals, the second tone of each stimulus and the first tone of each sung response can be considered to be realized intervals. The lack of melodic context may even provide a more direct test of the model's principles than did Experiment 1, 2, and 3, with their more complex melodic contexts. Access to the data sets from these two studies (J. Carlsen, personal communication, November, 1990; A. Unyk, personal communication, November, 1990) provided an opportunity to test the I-R model with data derived from a different method and with four different groups of listeners (individual listener responses were not available).

Instead of multiple regression analyses, multinomial log-linear analyses (appropriate for frequency data) were performed (D. Madigan, personal communication, June, 1991). The predictor variables for the five core principles of the I-R model and the three from the revised model were coded as in Experiments 1, 2, and 3. Effects of tonality were controlled by including a covariate called TONALITY, coded assuming that the musical key of each two-tone stimulus was perceived to be the major key of the first tone in the stimulus interval. The value of the predictor variable assigned to each of the 12 tones from the chromatic scale was the corresponding value from the tonal hierarchy (Krumhansl & Kessler, 1982) of that key. Several other methods of defining key were explored, including the minor key of the first tone and the major and minor keys of the second tone. None of these alternative methods accounted for as much variance in listeners' responses as the method originally selected.

13. Results of reanalysis

Data from responses to stimulus intervals of 6 and 12 semitones were excluded from the analyses because of Narmour's (1990, 1992) contention that the tritone (6 semitones) is a threshold value (i.e., neither small nor large), and that the octave (12 semitones) is an atypical large interval because of octave equivalence. Thus, the data were reanalyzed for 525 possible responses: 275 responses to small implicative intervals (25 possible responses for 0 semitones and for each of the five intervals from 1 semitone to 5 semitones in both ascending and descending forms), and 250 responses to large implicative intervals (25 possible responses for each of the five intervals from 7 semitones to 11 semitones in both ascending and descending forms).

13.1. Agreement among listener groups

Agreement among listener groups was high. The rank-order (Spearman) inter-culture correlations in Carlsen's (1981) data were: Germany–Hungary,

$r_s = .870$; Germany–USA, $r_s = .865$; and Hungary–USA, $r_s = .852$ ($N = 525$, $ps < .0001$). The data from Unyk and Carlsen's (1987) American listeners were also highly correlated with Carlsen's German listeners, $r_s = .828$, Hungarian listeners, $r_s = .776$, and American listeners, $r_s = .786$ ($N = 625$, $ps < .0001$).

13.2. I-R model

A preliminary analysis examined correlations between pairwise combinations of predictor variables for the 525 response cells in the present data matrix (corrected for 15 tests); these are provided in Table 7 (upper portion). As in Experiments 1, 2, and 3, INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE were significantly intercorrelated. Due to the large sample size and the corresponding increase in statistical power, three additional pairwise associations were small but statistically significant: REGISTRAL RETURN was significantly correlated with both CLOSURE and TONALITY, and REGISTRAL DIRECTION and INTERVALLIC DIFFERENCE were significantly correlated.

Log-linear analyses were used to test whether the principles of the I-R models were significant predictors of the response patterns from Carlsen's (1981) and Unyk and Carlsen's (1987) listeners. The model was fit separ-

Table 7
Correlations between predictor variables in Carlsen (1981)/Unyk and Carlsen (1987) reanalysis ($N = 525$)

I-R model					
	Intervallic difference	Registral return	Proximity	Closure	Tonality
Registral direction	.143 [†]	.033	.009	.012	.004
Intervallic difference		.036	.627***	.323***	.010
Registral return			.027	.124*	.160**
Proximity				.388***	.003
Closure					-.013

Revised model			
	Registral return (revised)	Proximity (revised)	Tonality
Registral direction (revised)	.336***	-.048	-.004
Registral return (revised)		.003	.008
Proximity (revised)			-.002

[†] $p < .05$; ** $p < .005$; *** $p < .0001$.

ately to the data from each listener group, and the results are summarized in Table 8 (upper portion). For each group, the model significantly reduced the deviance (i.e., the unexplained variation, distributed chi-square) from the independence model ($ps < .0001$), which is based solely on the grand mean of the number of responses per cell.

Each predictor variable was tested by removing it from the model and examining the significance of the resulting increase in deviance. The large number of responses cells (525) meant a substantial increase in statistical

Table 8

Log-linear results for Carlsen (1981)/Unyk and Carlsen (1987) reanalysis, including the proportion of deviance explained by the model (R^{2*}), and, for each predictor variable, the odds ratio (O.R.) and the proportion of deviance uniquely explained by the predictor (sr^{2*}). All $ps < .0001$

	Carlsen (1981)			Unyk and Carlsen (1987)
	American sample	German sample	Hungarian sample	American sample
<i>I-R model</i>	$R^{2*} = .540$	$R^{2*} = .598$	$R^{2*} = .541$	$R^{2*} = .567$
Registral direction	O.R. = 2.25 $sr^{2*} = .053$	O.R. = 1.65 $sr^{2*} = .021$	O.R. = 1.69 $sr^{2*} = .022$	O.R. = 2.19 $sr^{2*} = .049$
Intervallic difference	O.R. = 1.58 $sr^{2*} = .009$	O.R. = 1.87 $sr^{2*} = .015$	O.R. = 2.33 $sr^{2*} = .028$	O.R. = 1.90 $sr^{2*} = .016$
Registral return	O.R. = 1.29 $sr^{2*} = .014$	O.R. = 1.21 $sr^{2*} = .007$	O.R. = 1.23 $sr^{2*} = .008$	O.R. = 1.47 $sr^{2*} = .035$
Proximity	O.R. = 1.58 $sr^{2*} = .203$	O.R. = 1.63 $sr^{2*} = .227$	O.R. = 1.54 $sr^{2*} = .177$	O.R. = 1.53 $sr^{2*} = .164$
Closure	O.R. = .802 $sr^{2*} = .005$	O.R. = .775 $sr^{2*} = .007$	O.R. = .752 $sr^{2*} = .008$	O.R. = .785 $sr^{2*} = .006$
Tonality	O.R. = 1.21 $sr^{2*} = .023$	O.R. = 1.32 $sr^{2*} = .038$	O.R. = 1.21 $sr^{2*} = .024$	O.R. = 1.34 $sr^{2*} = .055$
<i>Revised model</i>	$R^{2*} = .513$	$R^{2*} = .580$	$R^{2*} = .516$	$R^{2*} = .530$
Registral direction (revised)	O.R. = 1.48 $sr^{2*} = .024$	O.R. = 1.20 $sr^{2*} = .005$	O.R. = 1.20 $sr^{2*} = .006$	O.R. = 1.45 $sr^{2*} = .021$
Registral return (revised)	O.R. = 1.35 $sr^{2*} = .005$	O.R. = 1.21 $sr^{2*} = .002$	O.R. = 1.32 $sr^{2*} = .004$	O.R. = 1.43 $sr^{2*} = .007$
Proximity (revised)	O.R. = .711 $sr^{2*} = .425$	O.R. = .680 $sr^{2*} = .501$	O.R. = .698 $sr^{2*} = .454$	O.R. = .723 $sr^{2*} = .384$
Tonality	O.R. = 1.28 $sr^{2*} = .046$	O.R. = 1.34 $sr^{2*} = .063$	O.R. = 1.28 $sr^{2*} = .044$	O.R. = 1.47 $sr^{2*} = .112$

power to detect differences in cell frequencies as a function of the unique contributions of each predictor variable. Hence, all six predictors were statistically significant for each group of respondents, and all odds ratios were significantly different from 1 (see Table 8, upper portion). For each predictor variable, odds ratios were formed by raising e to the power of the corresponding coefficient from the model. For principles coded as dummy variables, an odds ratio greater than 1 means that the odds that responses satisfied the principle were greater than the odds that responses violated the principle (with all other predictor variables held constant). For example, for Carlsen's (1981) American sample, the odds of a sung response satisfying REGISTRAL DIRECTION were 2.25 times greater than the odds of a response violating the principle. For graded predictor variables, each unit increase in the predictor is accompanied by a multiplicative increase in the odds of a particular response. For example, for Carlsen's (1981) American sample, the odds of a response with PROXIMITY coded as 1 were 1.58 times greater than the odds of a response coded as 0, the odds of a response coded as 2 were 1.58 times greater than the odds of a response coded as 1, and so on. An odds ratio less than 1 was found for CLOSURE (i.e., negative model coefficient) because of listeners' tendency to begin their sung responses with tones that did *not* cause closure.

For each group of respondents, the I-R model explained about 55% of the scaled deviance in the fit of the model to the data (see Table 8, upper portion). The percentage of deviance uniquely explained by each predictor variable is also provided in Table 8 (upper portion). The sum of these unique contributions explained approximately 30% of the deviation in the data. Thus, as in Experiments 1, 2, and 3, a substantial portion of the variation (deviance) in these data (approximately 25%) was explained redundantly by the I-R model.

13.3. Revised model

Results from a preliminary analysis, which examined correlations between pairwise combinations of predictor variables (corrected for six tests), were identical to those from Experiments 1, 2, and 3 (see Table 7, lower portion). That is, REGISTRAL DIRECTION (REVISED) and REGISTRAL RETURN (REVISED) were significantly correlated.

The revised model was fit separately to the data from each group of respondents. The results are summarized in Table 8 (lower portion). The revised model significantly reduced the deviation in the data for each group, and all four predictor variables made a significant unique contribution to the fit of the model in each analysis ($ps < .0001$, corrected for four tests).

The revised model accounted for approximately 53% of the deviance for each group of respondents, slightly less than that explained by the original model (about 3% less for each group). For each group, however, the revised model did not significantly differ from the original model in its ability to

predict cell frequencies. Unlike the original model, moreover, the redundancy of the revised model was minimal. The deviance accounted for by the unique contributions of the individual predictor variables revealed that, for each group of respondents, the proportion of deviance explained redundantly was less than 2%.

14. Discussion of reanalysis

The data of Carlsen (1981) and Unyk and Carlsen (1987) provided convergent evidence for the findings of Experiments 1, 2, and 3. Despite the difference in experimental task (sung continuations as opposed to the perceptual judgments of Experiments 1, 2, and 3), the data were well-fit by the predictor variables of both versions of the model (the original I-R model and the revised model). This convergence between perception and production data, like that reported by Schmuckler (1989, 1990), implies that the two measures are tapping the same system of musical expectancy. The only important difference between the production and perception data reported here is in the effect of CLOSURE as coded in the original model. Whereas tones that did not create closure were sung more frequently in the production task, tones that created closure were judged as better continuations in the perception task. This difference might be attributable to the very brief (two-tone) contexts provided by Carlsen (1981) and Unyk and Carlsen (1987) in their tasks, which may have generated expectancies for extended continuations. In contrast, the more extended contexts of the perceptual tasks of Experiments 1–3 might have created expectancies for more immediate closure.

The data from German, Hungarian, and American listeners (Carlsen, 1981) were strongly intercorrelated. Both models provided a highly significant fit to the data from each listener group, and the unique contributions of each of the predictor variables were consistent across groups. Moreover, the data from Unyk and Calsen's (1987) American listeners were strongly correlated with the data from each of Carlsen's German, Hungarian, and American listener groups, with similar results in the fit of the models. Whereas Carlsen (1981) emphasized the *differences* between his listener groups, the present reanalysis highlights the cross-cultural *similarities*. This discrepancy can be explained largely by differences in the level of analysis. Carlsen's (1981) analysis focused on individual response tones, whereas the present reanalysis focused on sets of tones. Carlsen considered a response of an upward major third (4 semitones) to an upward major third stimulus to be different from a response of an upward minor third (3 semitones). By contrast, both responses were similar in terms of the predictors of the I-R model or the revised model. Thus, by broadening the focus of the analysis to sets of tones, the models revealed cross-cultural similarities that might otherwise be overlooked.

Responses from all groups of listeners indicated that they interpreted the two-tone stimulus intervals as suggesting a particular key, namely the major key of the first stimulus tone, despite the impoverished melodic context. Indeed, TONALITY consistently explained a greater unique proportion of variation in these data than did the corresponding covariates in Experiments 1 and 3. This finding is likely due to the fact that “out-of-key” responses were considered in the present reanalysis, but not in Experiments 1 and 3.

A comparison of the I-R model and its revised counterpart revealed results similar to those of Experiments 1, 2, and 3. Because the revised model explained the data as well as the original but in a more parsimonious manner, it would seem to be superior for characterizing melodic expectancies in a vocal production context as well as in the perceptual rating contexts of Experiments 1, 2, and 3.

GENERAL DISCUSSION

The present report examined the claims of Narmour’s (1990, 1992) implication-realization model of melody. According to Narmour, a small number of universal psychological principles affect listeners’ expectancies of how a melody will continue. The model delineates these principles in terms of the parameters of interval size and pitch direction, which depend specifically on the size of an unclosed interval (i.e., an implicative interval) in a melody. The principles, although expressed in musical terms, are presumed to arise from general psychological processes. As a result, they would operate independently of the musical style of a melody and the musical experience of a listener.

The model successfully predicted: (1) listeners’ ratings of test tones following melodic fragments in a familiar Western musical style (tonal melodies, Experiment 1) and an unfamiliar Western style (atonal melodies, Experiment 2), regardless of listeners’ musical training, (2) American and Chinese listeners’ ratings of Chinese melodies (Experiment 3), and (3) American, German, and Hungarian listeners’ sung continuations of two-tone melodic intervals (Reanalysis). Overall the model was remarkably consistent in its ability to predict response patterns.

In each instance, however, the model was shown to be overspecified. A revised, simplified version of the model was equally successful in predicting responses across differences in stimulus materials, listener groups, individual listeners, and experimental methods. Moreover, the revised model, with fewer predictor variables, was consistently better in predicting response patterns of individual listeners (Experiments 1, 2, and 3).

Three of the five core principles of the I-R model were significantly intercorrelated across all three experiments and the reanalysis, resulting in explanatory redundancy. The revised model reduced this redundancy by eliminating two of the original intercorrelated principles and modifying the

third. The remaining two (non-intercorrelated) principles of the original model were also modified.

The intercorrelations between principles of the I-R model highlight their conceptual overlap. For example, the principle of intervallic difference states that a small implicative interval implies another small interval, whereas a large implicative interval implies a smaller interval. Thus, regardless of whether an implicative interval is small or large, the proposed implication is that the next tone of the melody will be proximate to the second tone of the implicative interval. In effect, the principle of intervallic difference is a principle of proximity, resulting in overlap with the proximity principle itself. The principle of closure has similar conceptual redundancy. One factor contributing to melodic closure involves a relatively smaller interval (and, hence, a relatively proximate next tone) following a larger interval. A separate problem is that the proximity principle, as originally quantified, assumes “proximate” to be equivalent to “small” (i.e., less than six semitones), such that all “large” intervals are equally non-proximate. This assumption has not been empirically validated. The revised version of the proximity principle simply states that a tone that is more proximate is more expected or implied than tones that are less proximate.

The revised model is not entirely free of conceptual overlap between its principles. The significant correlation between the revised principles of registral direction and registral return reflects the fact that for large implicative intervals, instances that satisfy the revised registral return principle involve a change of direction, thereby satisfying the revised registral direction principle as well. One solution would be to limit registral return so that it applied only to small intervals. In exploratory analyses, such a principle was derived and tested; it resulted in a loss of predictive power and was significantly correlated with the revised proximity predictor. In the future, it may be possible to devise a model of melodic expectancy that consists solely of orthogonal or nearly orthogonal predictor variables. At present, however, the revised model contains some collinearity between predictors. Although predictive accuracy is improved when predictors are uncorrelated, partial redundancy is the “standard configuration” in multiple regression models (Darlington, 1990, p. 150).

The results of the present study raise doubts about the psychological validity of the I-R model’s 12 basic melodic structures, which are based solely on adherence to (and violations of) intervallic difference and registral direction. Neither of these principles was retained in its original formulation in the revised model. Specifically, registral direction was modified so that it applied only to large intervals, whereas intervallic difference was omitted altogether. Thus, neither principle, as originally formulated, was necessary for explaining the data reported in the present study. This finding may stem from the fact that both of these principles are based on an arbitrary distinction between small and large implicative intervals (<6 semitones = small, >6 semitones = large) that is unjustified on empirical or theoretical

grounds. Moreover, the principle of intervallic difference suffers from an additional assumption (also without empirical validation) that similarity in interval size differs as a function of registral direction.

The results also indicated that the melodic structure known as registral return requires modification in terms of its specific formulation. In all of the data sets reported here, registral return was a better predictor when coded as a dummy variable, implying that the claim of a graded effect for registral return is untenable. Perhaps the observed threshold function is related to the claim that tone sequences exhibiting registral return are melodic archetypes. As such, sequences may either evoke archetypes or not, with few sequences perceived as “somewhat” archetypal.

According to the I-R model (Narmour, 1990, 1992), principles governing melodic expectancy are innately specified and therefore universal. The claim of “innateness” is not a requirement for universality, because universal processes may be universally learned. Although the present report is consistent with universality, the most that can be said is that the principles are operative under a wide range of musical circumstances. Moreover, the generality of the perceptual principles does not provide evidence that such principles are hard-wired, as Narmour (1990, 1992) claims.

How might we understand the psychological significance of universal principles governing melodic expectations? The principles, particularly those from the revised model, can be seen as closely related to, and possibly derived from, principles of perceptual organization that function in audition as well as vision. For example, the strongest predictor for all data sets was the revised principle of proximity. Proximity is a robust predictor of grouping in vision (Koffka, 1935; Kohler, 1947) and audition (Bregman, 1990) in general, so it is not surprising that it would influence the perception of complex auditory patterns, such as melodies (Dowling, 1973), in particular. The predominance of small intervals (proximate tones) in melodies across musical cultures (Dowling & Harwood, 1986) provides further support for the idea that proximity is a universal principle influencing the perception and cognition of melodies. The importance of proximity could also stem, in part, from vocal production limitations, because small intervals are easier than large intervals to sing (Bregman, 1990).

Because the principle of registral return describes a reversal of melodic direction and a return to an earlier pitch range, it describes a melodic archetype exhibiting proximate pitch relations between discontinuous tones. Registral return also describes a pitch pattern that is symmetric or approximately symmetric about a point in time. Symmetry, like proximity, is an important factor in vision, facilitating perceptual processing because of redundant information (Garner, 1970, 1974). Symmetry may also be a basis of processing predispositions with auditory stimuli (Schellenberg & Trehub, 1994).

The revised principle of registral direction, which states that a large

interval implies a change of melodic direction, may be a by-product of the proximity principle. A large interval in a melody violates the principle of proximity because it occurs between tones that are relatively far apart. Such non-proximate intervals may cause a lack of coherence in a melody at the tone-to-tone level (Bregman, 1990). If a melody reverses direction after the large interval, however, the gap between tones of the large interval is more likely to be filled in, which would increase the coherence of the melody as a whole (see Meyer, 1973). It is also conceivable that the registral direction principle stems from production constraints, reflecting limitations of the human vocal range. A large interval is more likely than a small interval to approach the limits of a singer's vocal range, making a change in pitch direction relatively more likely after a large interval. Thus, the expectancy for a reversal of pitch direction after a large interval could be a by-product of production limitations or exposure to melodies that reflect these limitations. The revised version of the registral direction principle retains the untested assumption that large intervals are those seven semitones or larger. Future research could explore the validity of a psychological threshold for distinguishing large intervals from small intervals.

The importance of proximity throughout the present report implies that listeners were applying general perceptual principles to a variety of stimulus contexts and experimental tasks. Indeed, one might argue that listeners' emphasis on proximate relations suggests that they were ignoring the musical aspects of the stimuli. Several aspects of the results provide evidence against such an interpretation. For example, the influence of tonality provides a parsimonious explanation of differences between data sets. In tonal contexts (Experiments 1 and 3), listeners's response patterns were far more regular than in atonal contexts (Experiment 2). Moreover, the revised model's account of response patterns to Carlsen's (1981; Unyk & Carlsen, 1987) relatively impoverished (two-tone) stimuli revealed that the tonality covariate made the second largest unique contribution to the model (after proximity). Thus, when considered as a whole, the data in the present report indicate that tonality had a strong effect on listeners' patterns of responding. The stimuli would not have evoked tonal schemas if listeners had perceived them to be non-musical. Narmour's (1990, 1992) description of the I-R model's principles as hard-wired implies, however, that these principles should be operative even in non-musical contexts. Future research could test this claim using the principles from the original model as well as those from the revised model.

The principles of the revised model proposed in the present report are similar to the original principles of the I-R model in their precision and specificity to musical pitch. The revised principles have the advantage, however, of more simply reflecting the application of general processes of perceptual organization (e.g., proximity, symmetry, familiarity) to music. The simplified proximity principle (i.e., *more proximate = more implied*)

was shown to be particularly robust. Moreover, the revised model, with its smaller number of principles, is simpler than the original I-R model, making it more likely to withstand future tests of its universality.

Acknowledgements

Most of the data reported here were collected under the supervision of Carol L. Krumhansl while the author was a graduate student at Cornell University. Parts of this article are excerpted from an unpublished paper by Carol L. Krumhansl and E. Glenn Schellenberg entitled “Melodic expectancy: psychological tests of the implication–realization model”. The research was supported by awards to E.G. Schellenberg from the Natural Sciences and Engineering Research Council of Canada, the University of Toronto, the Psychology Department at Cornell University, and the University of Windsor, and by a Grant-in-Aid of Research from Sigma Xi, the Scientific Research Society. I thank Sandra Trehub, Laurel Trainor, and two anonymous reviewers for their helpful comments on earlier versions of the manuscript, Bill Thompson for collegial discussions about Narmour’s theory, James Carlsen and Anna Unyk for providing their data for reanalysis, Dick Darlington, Gary Churchill, and David Madigan for statistical assistance, Martin Hatch and Edward Murray for their expertise in Chinese music and Webern *Lieder*, respectively, and Adrian Roberts for writing the computer program for Experiments 1, 2, and 3. I am especially grateful to Eugene Narmour for extensive consultation about the implication-realization model.

Appendix A: Averaged rating for each test tone in Experiment 1 (20 listeners)

Fragment 1			Fragment 2			Fragment 3			Fragment 4			Fragment 5			Fragment 6			Fragment 7			Fragment 8		
Test tone	Avg. rating		Test tone	Avg. rating		Test tone	Avg. rating		Test tone	Avg. rating		Test tone	Avg. rating		Test tone	Avg. rating		Test tone	Avg. rating		Test tone	Avg. rating	
G ₅	2.15		C ₆	1.55		B ₅	1.85		F ₅	2.65		F ₅	4.45		A ₅	1.75		E ₅	3.70		C ₆	2.35	
F# ₅	1.85		Bb ₅	1.70		A ₅	1.65		E ₅	1.80		Eb ₅	4.40		G ₅	2.05		D ₅	4.10		Bb ₅	1.80	
E ₅	2.75		Ab ₅	2.00		G ₅	3.55		D ₅	4.15		D ₅	5.40		F ₅	2.40		C ₅	5.15		A ₅	2.40	
D ₅	4.80		Gb ₅	2.35		F# ₅	2.45		C ₅	3.60		C ₅	5.45		E ₅	2.65		B ₅	4.80		G ₅	2.85	
C# ₅	3.25		F ₅	3.40		E ₅	5.00		Bb ₅	4.95		Bb ₅	5.55		D ₅	3.35		A ₅	4.60		F ₅	3.85	
B ₅	4.55		Eb ₅	4.10		D ₅	5.40		A ₅	6.05		A ₅	4.15		C ₅	5.00		G ₅	4.95		E ₅	3.80	
A ₅	5.15		Db ₅	5.70		C ₅	5.00		G ₅	6.15		G ₅	5.80		B ₅	3.80		F# ₅	4.80		D ₅	5.25	
G ₄	5.50		C ₅	5.75		B ₅	5.85		F ₅	5.20		F ₅	5.35		A ₅	5.35		E ₅	4.35		C ₅	5.20	
F# ₄	5.95		Bb ₄	5.80		A ₅	5.85		E ₅	6.00		Eb ₄	3.60		G ₄	6.40		D ₅	4.25		Bb ₄	4.45	
E ₄	3.50		Ab ₄	5.30		G ₄	5.20		D ₄	5.35		D ₄	4.95		F ₄	4.55		C ₄	3.05		A ₄	5.10	
D ₄	3.20		Gb ₄	2.85		F# ₄	2.75		C ₄	3.80		C ₄	2.85		E ₄	4.45		B ₅	2.90		G ₄	4.60	
C# ₄	2.15		F ₄	5.25		E ₄	3.15		Bb ₅	3.25		Bb ₅	3.15		D ₄	4.20		A ₅	2.20		F ₄	4.10	
B ₅	2.15		Eb ₄	3.15		D ₄	3.15		A ₅	3.35		A ₅	1.85		C ₁	4.65		G ₅	2.05		E ₄	4.15	
A ₅	1.75		Db ₄	2.50		C ₁	1.55		G ₅	2.25		G ₅	2.15		B ₅	3.15		F# ₅	1.60		D ₄	4.95	
G ₅	1.45		C ₄	2.80		B ₅	2.20		F ₅	2.30		F ₅	2.70		A ₅	2.65		E ₅	2.00		C ₄	4.40	

Appendix B: Averaged rating for each test tone in Experiment 2 (22 listeners, 4 outliers excluded)

Fragment 1		Fragment 2		Fragment 3		Fragment 4		Fragment 5		Fragment 6		Fragment 7		Fragment 8	
Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating
F# ₅	2.27	F ₅	4.18	A ₅	3.77	G# ₅	2.36	D# ₅	5.23	A ₅	4.32	C ₅	4.45	G# ₆	2.45
F ₅	1.86	E ₅	3.00	G# ₅	2.82	G# ₅	1.82	D ₅	4.86	G# ₅	4.27	B ₄	4.23	G ₆	2.18
E ₅	2.05	D# ₅	4.05	G ₅	3.32	F# ₅	2.09	C# ₅	5.18	G ₅	4.09	A# ₄	4.27	F# ₆	2.05
D# ₅	3.05	D ₅	4.00	F# ₅	4.45	F ₅	2.45	C ₅	4.68	F# ₅	3.95	A ₄	4.82	F ₆	2.77
D ₅	3.50	C# ₅	5.27	F ₅	4.14	E ₅	3.14	B ₄	5.18	F ₅	4.36	G# ₄	4.73	E ₆	2.45
C# ₅	4.14	C ₅	5.73	E ₅	4.55	D# ₅	3.41	A# ₄	5.23	E ₅	4.36	G ₄	5.23	D# ₆	3.64
C ₅	3.55	B ₄	5.00	D# ₅	4.27	D ₅	3.00	A ₄	5.64	D# ₅	4.41	F# ₄	5.36	D ₆	2.95
B ₄	4.50	A# ₄	5.82	D ₅	5.64	C# ₅	4.23	G# ₄	5.00	D ₅	4.73	F ₄	5.64	C# ₆	4.14
A# ₄	4.95	A ₄	5.32	C# ₅	5.55	C ₅	4.36	G ₄	5.45	C# ₅	4.23	E ₄	5.27	C ₆	4.36
A ₄	5.59	G# ₄	6.00	C ₅	5.23	B ₄	5.59	F# ₄	5.73	C ₅	4.18	D# ₄	5.32	B ₅	4.59
G# ₄	5.45	G ₄	5.59	B ₄	5.50	A# ₄	5.41	F ₄	4.91	B ₄	4.82	D ₄	4.86	A# ₅	5.00
G ₄	5.59	F# ₄	4.59	A# ₄	5.00	A ₄	5.18	E ₃	5.86	A# ₄	4.73	C# ₄	5.64	A ₅	4.55
F# ₄	4.91	F ₄	4.73	A ₄	4.86	G# ₄	4.55	D# ₄	3.91	A ₄	4.23	C ₄	4.73	G# ₅	4.50
F ₄	4.91	E ₄	4.45	G# ₄	4.95	G ₄	6.05	D ₄	4.64	G# ₄	5.55	B ₃	5.23	G ₅	5.86
E ₄	4.95	D# ₄	2.73	G ₄	5.14	F# ₄	5.18	C# ₄	4.55	G ₄	4.73	A# ₃	3.86	F# ₅	4.91
D# ₄	4.95	D ₄	3.23	F# ₄	4.73	F ₄	4.95	C ₄	3.91	F# ₄	4.32	A ₃	3.23	F ₅	5.50
D ₄	5.36	C# ₄	2.86	F ₄	4.36	E ₄	4.50	B ₃	4.59	F ₄	4.91	G# ₃	3.32	E ₅	5.86
C# ₄	5.23	C ₄	2.82	E ₄	4.55	D# ₄	4.64	A# ₃	4.23	E ₄	4.86	G ₃	2.77	D# ₅	5.50
C ₄	4.68	B ₃	2.36	D# ₄	2.64	D ₄	3.77	A ₃	3.09	D# ₄	3.95	F# ₃	2.50	D ₅	4.45
B ₃	4.77	A# ₃	2.59	D ₄	3.64	C# ₄	4.05	G# ₃	2.86	D ₄	4.36	F ₃	2.95	C# ₅	4.82
A# ₃	4.14	A ₃	1.77	C# ₄	3.23	C ₄	2.82	G ₃	2.59	C# ₄	3.95	E ₃	2.36	C ₅	3.91
A ₃	3.68	G# ₃	2.05	C ₄	2.05	B ₃	3.09	F# ₃	2.36	C ₄	4.05	D# ₃	2.18	B ₄	4.73
G# ₃	2.68	G ₃	1.64	B ₃	1.82	A# ₃	2.36	F ₃	2.14	B ₃	4.09	D ₃	1.64	A# ₄	4.23
G ₃	2.68	F# ₃	1.41	A# ₃	1.59	A ₃	1.59	E ₃	2.41	A# ₃	3.82	C# ₃	2.32	A ₄	3.50
F# ₃	2.32	F ₃	2.27	A ₃	2.36	G# ₃	2.32	D# ₃	1.86	A ₃	3.77	C ₃	2.55	G# ₄	4.86

**Appendix C: Averaged rating for each test tone in Experiment 3
(16 listeners)**

Fragment 1		Fragment 2		Fragment 3		Fragment 4		Fragment 5		Fragment 6	
Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating
G ₅	1.75	F# ₅	1.50	G ₆	2.06	C ₆	1.38	A ₅	1.63	D ₆	1.81
F ₅	2.25	E ₅	1.94	F ₆	2.44	A ₆	1.88	F# ₅	2.06	C ₆	2.25
D ₅	3.81	D ₅	2.69	D ₆	2.63	G ₅	2.19	E ₅	3.31	Bb ₅	2.88
C ₅	5.00	B ₅	4.13	C ₆	3.75	F ₅	4.25	C# ₅	5.50	G ₅	3.50
A ₄	5.81	A ₄	5.38	A ₅	4.94	D ₅	5.94	B ₄	5.50	F ₅	4.69
G ₄	5.94	F# ₄	5.00	G ₅	5.50	C ₅	5.50	A ₄	5.75	D ₅	4.94
F ₄	5.88	E ₄	5.06	F ₅	6.13	A ₄	5.44	F# ₄	5.50	C ₅	6.13
D ₄	3.75	D ₄	5.13	D ₅	3.88	G ₄	5.25	E ₄	4.69	Bb ₄	4.56
C ₄	2.38	B ₄	5.38	C ₅	3.75	F ₄	4.31	C# ₄	3.31	G ₄	5.25
A ₃	1.81	A ₃	3.38	A ₄	2.75	D ₄	2.38	B ₃	2.38	F ₄	3.94
G ₃	1.50	F# ₃	2.25	G ₄	2.13	C ₄	2.00	A ₃	2.88	D ₄	3.13

Fragment 7		Fragment 8		Fragment 9		Fragment 10		Fragment 11		Fragment 12	
Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating	Test tone	Avg. rating
G# ₅	2.38	C ₆	2.38	D ₅	4.31	F# ₅	2.25	D ₅	3.81	C ₆	1.81
F# ₅	2.44	A ₅	1.94	B ₄	4.88	E ₅	2.38	C ₅	4.31	A ₅	1.56
E ₅	4.56	G ₅	2.75	A ₄	4.31	D ₅	3.56	A ₄	4.63	G ₅	2.38
C# ₅	5.63	E ₅	3.94	G ₄	5.88	B ₅	3.06	G ₄	5.00	F ₅	3.06
B ₅	5.75	D ₅	4.19	E ₄	5.25	A ₄	5.13	F ₄	5.81	D ₅	4.50
G# ₄	5.44	C ₅	4.69	D ₄	5.00	F# ₄	5.44	D ₄	4.25	C ₅	5.50
F# ₄	5.31	A ₄	5.88	B ₅	3.75	E ₄	5.81	C ₄	5.13	A ₄	5.63
E ₄	4.38	G ₄	4.88	A ₅	2.88	D ₄	5.31	A ₅	2.19	G ₄	5.81
C# ₄	2.63	E ₄	4.31	G ₅	2.88	B ₅	3.13	G ₅	2.19	F ₄	5.13
B ₄	2.56	D ₄	4.31	E ₅	2.88	A ₅	3.31	F ₅	2.63	D ₄	3.75
G# ₄	2.19	C ₄	3.50	D ₅	2.69	F# ₅	2.75	D ₅	1.81	C ₄	3.63

References

- Bartlett, J.C., & Dowling, W.J. (1988). Scale structure and similarity of melodies. *Music Perception*, 5, 285–314.
- Boltz, M., & Jones, M.R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, 18, 389–431.
- Boltz, M., Marshburn, E., Jones, M.R., & Johnson, W.W. (1985). Serial-pattern structure and temporal-order recognition. *Perception & Psychophysics*, 37, 209–217.
- Bregman, A.S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

- Carlsen, J.C. (1981). Some factors which influence melodic expectancy. *Psychomusicology*, 1, 12–29.
- 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.
- 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 musical rules. *Canadian Journal of Psychology*, 33, 148–157.
- Cutting, J.E., Bruno, N., Brady, N.P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, 121, 364–381.
- Darlington, R.B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Deutsch, D. (1978). Delayed pitch comparisons and the principle of proximity. *Perception & Psychophysics*, 23, 227–230.
- Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics*, 28, 381–389.
- Deutsch, D., & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review*, 88, 503–522.
- Dowling, W.J. (1973). The perception of interleaved melodies. *Cognitive Psychology*, 5, 322–337.
- Dowling, W.J., & Harwood, D.L. (1986). *Music cognition*. Orlando, FL: Academic Press.
- Garner, W.R. (1970). Good patterns have few alternatives. *American Scientist*, 58, 34–42.
- Garner, W.R. (1974). *The processing of information and structure*. Hillsdale, NJ: Erlbaum.
- Gilman, B.I. (1892). On some psychological aspects of the Chinese musical system II. *Philosophical Review*, 1, 154–178.
- Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, MA: MIT Press.
- Kessler, E.J., Hansen, C., & Shepard, R.N. (1984). Tonal schemata in the perception of music in Bali and the West. *Music Perception*, 2, 131–165.
- Koffka, K. (1935). *Principles of Gestalt psychology*. London: Routledge & Kegan Paul.
- Kohler, W. (1947). *Gestalt psychology: An introduction to new concepts of modern psychology*. New York: Liveright.
- Koon, N.K. (1979). The five pentatonic modes in Chinese folk music. *Chinese Music*, 2 (2), 10–13.
- Krumhansl, C.L. (1990). *Cognitive foundations of musical pitch*. New York: Oxford University Press.
- Krumhansl, C.L. (1991). Music psychology: tonal structures in perception and memory. *Annual Review of Psychology*, 42, 277–303.
- Krumhansl, C.L., & Keil, F.C. (1982). Acquisition of the hierarchy of tonal functions in music. *Memory and Cognition*, 10, 243–251.
- 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., Sandell, G.S., & Sergeant, D.C. (1987). The perception of tone hierarchies and mirror forms in twelve-tone serial music. *Music Perception*, 5, 31–78.
- Laloy, L. (1979). *Chinese music* (N. Karel, Trans.). Paris: Henri-Laurens. (Original work published 1909.)
- Meyer, L.B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.
- Meyer, L.B. (1973). *Explaining music*. Berkeley, CA: University of California Press.
- Morrongiello, B.A., & Roes, C.L. (1990). Developmental changes in children's perception of musical sequences: effects of musical training. *Developmental Psychology*, 26, 814–820.
- Morrongiello, B.A., Roes, C.L., & Donnelly, F. (1989). Children's perception of musical patterns: effects of musical instruction. *Music Perception*, 6, 447–462.

- Narmour, E. (1989). The “genetic code” of melody: cognitive structures generated by the implication-realization model. *Contemporary Music Review*, 4, 45–64.
- Narmour, E. (1990). *The analysis and cognition of basic melodic structures: the implication-realization model*. Chicago: University of Chicago Press.
- Narmour, E. (1992). *The analysis and cognition of melodic complexity: the implication-realization model*. Chicago: University of Chicago Press.
- Palmer, R. (Ed.) (1983). *Folk songs collected by Ralph Vaughan Williams*. London: Dent.
- People’s Music Publishing Company (1980). *Chung-kuo min kuo hsuan* [Chinese folk songs]. Beijing: Author.
- Schellenberg, E.G., & Trehub, S.E. (1994). Perceptual processing predispositions: data and speculation. In I. Deliège (Ed.), *Proceedings of the third international conference for music perception and cognition* (pp. 129–130). Liège, Belgium: ESCOM.
- Schmuckler, M.A. (1989). Expectation in music: investigation of melodic and harmonic processes. *Music Perception*, 7, 109–150.
- Schmuckler, M.A. (1990). The performance of global expectations. *Psychomusicology*, 9, 122–147.
- Sharp, C.J. (Ed.) (1920). *English folk songs* (Vols. 1–2, selected ed.). London: Novello.
- Shu-hsien, H. (1986). On the writing of polyphonic music. *Chinese Music*, 9 (1), 9–16.
- Simon, H.A., & Sumner, R.K. (1968). Pattern in music. In B. Kleinmuntz (Ed.), *Formal representation of human judgement* (pp. 219–250). New York: Wiley.
- Sin-yan, S. (1979). Foundations of the Chinese orchestra (1). *Chinese Music*, 2 (3), 32–36.
- Sin-yan, S. (1981). What makes Chinese music Chinese? *Chinese Music*, 4 (2), 23–37.
- Trainor, L.J., & Trehub, S.E. (1992). A comparison of infants’ and adults’ sensitivity to Western musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 394–402.
- Unyk, A.M., & Carlsen, J.C. (1987). The influence of expectancy on melodic perception. *Psychomusicology*, 7, 3–23.
- Webern, A. (1921). *Fünf lieder* [Five songs], *op. 3*. Vienna: Universal Edition.
- Webern, A. (1923). *Fünf lieder* [Five songs], *op. 4*. Vienna: Universal Edition.
- Webern, A. (1924). *Fünf geistliche lieder* [Five sacred songs], *op. 15*. Vienna: Universal Edition.
- Yung, B.N. (1980). China IV: theory. In S. Sadie (Ed.), *The new Grove dictionary of music and musicians* (Vol. 4, pp. 260–262). Washington, DC: Grove’s Dictionary of Music, Inc.