Expectancy in Melody: Tests of Children and Adults

E. Glenn Schellenberg
University of Toronto

Mayumi Adachi
Hokkaido University

Kelly T. Purdy
McGill University

Margaret C. McKinnon
University of Toronto

Melodic expectancies among children and adults were examined. In Experiment 1, adults, 11-year-olds, and 8-year-olds rated how well individual test tones continued fragments of melodies. In Experiment 2, 11-, 8-, and 5-year-olds sang continuations to 2-tone stimuli. Response patterns were analyzed using 2 models of melodic expectancy. Despite having fewer predictor variables, the 2-factor model (E. G. Schellenberg, 1997) equaled or surpassed the implication-realization model (E. Narmour, 1990) in predictive accuracy. Listeners of all ages expected the next tone in a melody to be proximate in pitch to the tone heard most recently. Older listeners also expected reversals of pitch direction, specifically for tones that changed direction after a disruption of proximity and for tones that formed symmetric patterns.

Expectancies give rise to feelings of surprise, disappointment, fear, and closure, allowing humans to experience breathless anticipation and haunting anxiety. Because they allow humans (and animals) to plan ahead, expectancies are adaptive cognitive processes. Indeed, Dennett (1991) considered expectancies to be central to consciousness, referring to brains as "anticipation machines." The centrality of expectancies in human cognition (e.g., Dowling, 1990) is highlighted by claims that "the concept of 'expectancy' forms the basis for virtually all behavior" (Olson, Roese, & Zanna, 1996, p. 211). In the present investigation, we examined expectancies that are formed when listening to music. Musical expectancies are important because (a) listening to music is a universal behavior (e.g., Trehub, 2000), (b) emotion and meaning conveyed by music are thought to depend on whether expectancies are fulfilled or denied (Meyer, 1956), and (c) an improved understanding of musical expectancies could improve our understanding of expectancies in other domains.

We define *expectancy* as anticipation of an event based on its probability of occurring (Chaplin, 1985). Expectancies are implicit or explicit hypotheses about the future, which can be determined either by learned associations (nurture) or by cognitive predispositions (nature). Expectancies are categorized on the basis of the source of the expectancy (i.e., a stimulus or a behavior) and whether the expected consequence is a particular *response* or another *stimulus* (Maddux, 1999). A classic example of a response expectancy is the placebo effect. Learned associations between an environmental event (e.g., taking a pill, drinking coffee) and a particular response (e.g., relief from pain, stimulation) can be so strong that some people respond to inert substances (e.g., a sugar pill, decaffeinated coffee) much as they do to potent stimuli (e.g., prescription painkillers, espresso; Kirsh, 1999). In the musical domain, a familiar piece (e.g., a song played at a graduation, wedding, or funeral) can have a strong association with a particular emotion experienced in the past, leading to an expectancy to respond similarly when listening to the same piece in the future.

Stimulus expectancies refer to situations in which one stimulus is expected to be followed by another stimulus or by a particular environmental event. For example, the smell of freshly baked bread can lead one to expect that a bakery is nearby. Listening to music typically involves stimulus rather than response expectancies. Stimulus expectancies in the musical domain are further delineated into *veridical* and *schematic* expectancies (Bharucha, 1994). *Veridical expectancies* refer to listeners' anticipation of specific musical events in familiar musical pieces or tunes (e.g., for a saxophone solo in a particular pop song, for the next note in a familiar melody). Such expectancies give rise to *musical restoration effects,* which are evident when listeners fail to notice that a tone from a familiar melody has been replaced by a burst of noise (DeWitt & Samuel, 1990). Veridical expectancies let listeners focus their attention toward the point in time when the next tone in a familiar melody will occur (Jones, 1976), allowing for identification of the melody when it is presented in the midst of distractor tones (Dowling, 1973; Dowling, Lung, & Herrbold, 1987).

By contrast, schematic expectancies result from processing biases that listeners bring to the musical experience. Accordingly, listeners can have schematic expectancies when listening to a piece

for the first time. In many instances, such biases are a consequence of exposure to a particular style of music (e.g., Western tonal music). Among listeners familiar with Western music, a dominant-seventh (unresolved) chord creates an expectancy for a tonic (resolved) chord (e.g., the last two chords of the first movement of Beethoven's Fifth Symphony), even for those with no explicit knowledge of "tonic" or "dominant." Expectancies for tonic chords are evident in relatively impoverished contexts, such as when listeners hear a short sequence of chords belonging to a musical key (Bigand, Madurell, Tillmann, & Pineau, 1999; Bigand & Pineau, 1997; Tillmann, Bigand, & Pineau, 1998) or when a dominant chord is played in isolation (Bharucha & Stoeckig, 1986) for a very brief duration (i.e., 50 ms; Tekman & Bharucha, 1992).

But biases in music listening—and the schematic expectancies that result—could also be influenced by cognitive and perceptual predispositions (e.g., Jones, 1990; Thompson & Schellenberg, 2002; Trehub, 2000), including those specified by gestalt grouping principles. In the visual domain, the *violation-of-expectancy* paradigm is used to test young infants' perception of objects and events. Extended looking times are recorded for visual displays that depict violations of general perceptual principles or laws of physics, presumably because infants find these violations at odds with their innate or rapidly learned knowledge of the physical world (e.g., Baillargeon, 1999; Spelke, 1994). General predispositions could apply similarly to music listening and musical expectancies.

Our present focus was on stimulus expectancies that are formed when listening to melodies. *Melody* refers to a sequence of tones, and a melodic *interval*[1] refers to the distance in pitch between two successive tones (a portion of a melody). *Harmony* refers to simultaneously sounded tones (e.g., chords or chord progressions). Our study objectives comprised two major goals. Our first goal was to examine how melodic expectancies can be best explained. We compared two existing models of melodic expectancies on the basis of their *simplicity, scope,* and *selectivity* (Cutting, Bruno, Brady, & Moore, 1992). Following one of the basic rules of science, a simpler model is preferable to a more complex model. *Simplicity* can be measured by the number of freely varying parameters in a model (e.g., the number of predictor variables and interaction terms). Another feature of a good model is its ability to generalize across contexts (e.g., stimuli and experimental methods), or its *scope*. Obviously, a model should not be so general that it explains unrelated phenomena. One can test whether a model's scope is too broad by examining whether it predicts random data successfully. Finally, *selectivity* refers to a model's ability to explain data it should explain (i.e., patterned data in a relevant domain) better than random data.

Our second goal was to examine how melodic expectancies change over development. To this end, we tested participants who varied in age: adults and children of two different ages in Experiment 1, and children of three different ages in Experiment 2. In addition to the criteria outlined by Cutting et al. (1992), a good model of melodic expectancies should be able to describe, in a systematic manner, how such expectancies change over development. A third, ancillary goal was to investigate whether cognition and cognitive development in this area are music specific or reflective of general processes relevant to perception and cognition in other domains and modalities.

Models of Melodic Expectancy

We compared the explanatory power of two models of melodic expectancy: the *implication-realization* (I-R) model (Narmour, 1990) and the revised and simplified *two-factor* model (Schellenberg, 1997). Narmour, a music theorist, developed the I-R model as an explanatory framework for describing melodic expectancies. His focus is on schematic stimulus expectancies, or on how tones in an *unfamiliar* melody imply subsequent tones. The I-R model states explicitly that schematic expectancies are a function of learned factors acting in conjunction with innate factors.[2] Obvious differences between musical styles both within (e.g., jazz, pop, and classical) and across (e.g., Indian, Chinese, African, and European) cultures provide ample evidence that exposure to music and learning guide the formation of expectancies. Familiarity with style-specific musical structures (e.g., popular songs, typical chord progressions) gives rise to learned expectancies (e.g., Andrews & Dowling, 1991; Dowling, 2001). By contrast, the proposal of innate or culture-free principles of melodic expectancy is somewhat contentious and the focus of the present report.

How can one explain the diversity of musical styles if core aspects of music listening are innately guided or constrained by general cognitive–perceptual principles? Narmour's (1990) response to this apparent conundrum is that basic gestalt-based principles of perceptual organization are relevant to audition in general, and to music in particular, much as they are to vision. Bregman's (1990) research on auditory streaming reveals that sequentially presented tones are grouped on the basis of pitch proximity and timbral similarity, much like the way visual stimuli are grouped on the basis of spatial proximity or similarity in shape or color. According to the I-R model, these grouping principles are central to the formation of melodic expectancies. Because auditory signals unfold over time, time-based schemas of grouping in the auditory domain give rise to melodic expectancies. Listeners expect upcoming tones in a melody to be similar, proximate, and so on, to tones they have already heard.

The I-R model makes precise predictions about melodic expectancies that can be quantified and tested. Its predictions are most clearly specified at the tone-to-tone (lowest) level. Any melodic interval that is perceived as being "open" (incomplete sounding) is said to create an *implication* for the listener (see Figure 1). Because an open interval sounds unfinished, listeners expect that it will be followed by an additional tone or tones. Factors that contribute to closure (i.e., the opposite of openness) at the single-interval (two-tone) level include (a) when the second tone is longer in duration than the first tone (e.g., eighth note followed by quarter note), (b) when the second tone is more stable in the established musical key (e.g., *ti* followed by *do*), and (c) when the second tone falls on a beat with stronger metrical emphasis (e.g., at the first beat of a measure). At a more global level (i.e., two consecutive intervals, or three tones instead of two), a tone that reverses pitch direction is

---

[1] In the present article, the term *interval* refers to the distance in pitch between two tones and not to temporal intervals or to testing intervals (as in psychophysical methods).

[2] Narmour (1990) used the term *bottom-up* to describe perceptual predispositions among human listeners. Because the term typically refers to properties of the stimulus, we chose to use alternative terms such as *innate* or *hardwired.*
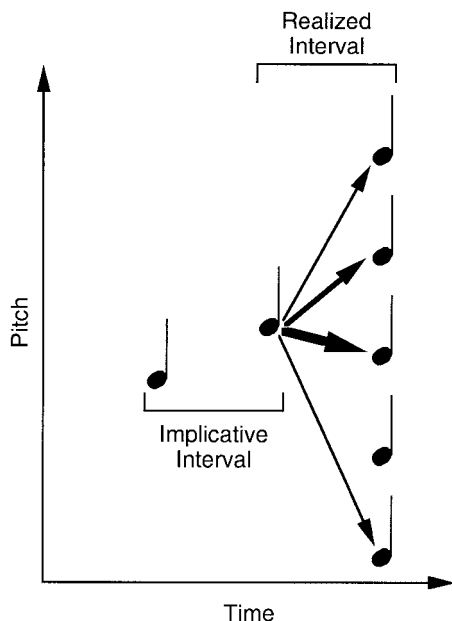
Figure 1. Schematic drawing of an implicative interval (between the first and second tones) followed by a realized interval (between the second and third tones). Any unclosed interval generates implications about the note to follow. The relative thickness of the arrows indicates that some tones are implied more than others.

said to cause closure, as does a large interval followed by a smaller interval (Narmour, 1990). An example of a three-note melody with both of these closural properties is the NBC chimes (an upward interval of 9 semitones, e.g., C_4 to A_4, followed by a downward interval of 4 semitones, e.g., to F_4)[3]; these properties contribute to making the melody sound like a complete pattern. The pattern also sounds complete because the final tone lasts for a longer duration than the first two tones and because the final tone is more stable in the implied musical key and meter.

Denial of some or all of the factors contributing to closure results in an interval that is partly or completely open, generating implications for the listener about the tone to follow. Typically, the strongest implications follow an interval that is completely open. Such implications are probability based rather than all-or-none. The set of possible tones that could follow an implicative interval includes some that are strongly implied, some that are moderately implied, others that are weakly implied, and still others that are not implied at all (see Figure 1). The term *realized interval* refers to the interval formed by the second tone of the implicative interval and the tone that follows. Both the implication and the realization describe relations between tones (i.e., intervals) rather than specific tones (i.e., absolute pitches or absolute durations). Thus, the I-R model describes how one melodic interval (an implicative interval) generates expectancies for a subsequent interval (a realized interval). This emphasis on musical relations is consistent with the way in which a melody is defined. A familiar tune such as "Yankee Doodle" can be performed quickly or slowly, or in a high or a low register, yet still be recognized provided the relations among tones conform to those of the song.

After careful reading of Narmour's (1990; see also Narmour 1989, 1992) theory and extensive personal communications,

Schellenberg (1996) quantified five principles that Narmour considers to be an accurate reflection of the hardwired components of the I-R model. Tables 1 and 2 provide examples of the quantified principles. Figure 2 illustrates eight different combinations of implicative and realized intervals in musical notation. The figure also indicates how each combination is quantified in the I-R model. Each of the five principles is discussed in turn.

Two principles—REGISTRAL DIRECTION and INTERVALLIC DIFFERENCE—form the core of the model. (Small uppercase letters are used in this article to designate quantified predictor variables and to make them distinct from concepts.) From these two principles, the I-R model's *basic melodic structures* are formed. A melody can be analyzed as a series of such structures arranged hierarchically, starting at the tone-to-tone level and becoming increasingly more abstract. For example, a melodic phrase could have seven structures at the lowest level but only one at the highest level. Both principles are quantified as dichotomous (dummy-coded) variables, specifying that one set of tones is more likely than another set to follow an implicative interval. In addition, both are a function of the size of the implicative interval. Narmour (1990) claimed that small implicative intervals, defined as 5 semitones (perfect fourths) or smaller, create expectancies for *similarity* in interval size and pitch direction. By contrast, large intervals, defined as 7 semitones (perfect fifths) or larger, generate expectancies for *change*. Intervals of 6 semitones (tritones) are considered ambiguous in terms of size (neither small nor large).

According to the principle of REGISTRAL DIRECTION, small implicative intervals lead to expectancies for similarity in pitch direction, specifically that the next (realized) interval will continue the direction of the melody (upward followed by upward, downward followed by downward, or lateral followed by lateral). For example, after a small, upward implicative interval of 2 semitones (e.g., C_4–D_4), another upward interval (e.g., D_4–E_4, D_4–D_5) is expected, but lateral or downward realized intervals (e.g., D_4–D_4, D_4–F_3) are unexpected (see Table 1, column 1, and Figure 2a–2d). By contrast, large intervals generate an expectancy for a change in direction, such as when a large, upward implicative interval of 9 semitones (e.g., C_4–A_4) creates expectancies for lateral or downward realized intervals (e.g., A_4–A_4, A_4–B_3) but not for another upward interval (e.g., A_4–B_4, A_4–A_5; see Table 2, column 1, and Figure 2e–2h).

The principle of INTERVALLIC DIFFERENCE states that small implicative intervals generate expectancies for realized intervals that are *similar* in size, whereas large implicative intervals create expectancies for *smaller* realized intervals. Similarity in size depends on whether pitch direction changes or remains constant. When implicative and realized intervals have the same direction, they are considered "similar" in size if they differ by 3 semitones or less (thus, "smaller" is 4 or more semitones smaller). When the realized interval changes direction, "similarity" is defined as a difference of 2 semitones or less (thus, "smaller" is 3 or more semitones smaller). For example, a small, upward implicative interval of 2 semitones (e.g., C_4–D_4) generates expectancies for similarly sized

---

[3] The subscript indicates the octave a particular tone is in. By convention, octaves are defined in relation to C. Middle C is C_4. The D above middle C is D_4, whereas the B below middle C is B_3. The C an octave lower or higher than middle C is C_3 or C_5, respectively.

Table 1

Quantification of the Principles From the Implication-Realization (I-R) and Two-Factor Models

Realized interval	I-R model					Two-factor model	
	Registral direction	Intervallic difference	Registral return	Proximity	Closure	Pitch proximity	Pitch reversal
$D_4–D_5$	1	0	0	0	0	12	0
$D_4–C_5^\#$	1	0	0	0	0	11	0
$D_4–C_5$	1	0	0	0	0	10	0
$D_4–B_4$	1	0	0	0	0	9	0
$D_4–A_4^\#$	1	0	0	0	0	8	0
$D_4–A_4$	1	0	0	0	0	7	0
$D_4–G_4^\#$	1	0	0	0	0	6	0
$D_4–G_4$	1	1	0	1	0	5	0
$D_4–F_4^\#$	1	1	0	2	0	4	0
$D_4–F_4$	1	1	0	3	0	3	0
$D_4–E_4$	1	1	0	4	0	2	0
$D_4–D_4^\#$	1	1	0	5	0	1	0
$D_4–D_4$	0	1	0	6	1	0	0
$D_4–C_4^\#$	0	1	2	5	1	1	1.5
$D_4–C_4$	0	1	3	4	1	2	1.5
$D_4–B_3$	0	1	2	3	1	3	1.5
$D_4–A_3^\#$	0	1	1	2	1	4	1.5
$D_4–A_3$	0	0	0	1	1	5	0
$D_4–G_3^\#$	0	0	0	0	1	6	0
$D_4–G_3$	0	0	0	0	1	7	0
$D_4–F_3^\#$	0	0	0	0	1	8	0
$D_4–F_3$	0	0	0	0	1	9	0
$D_4–E_3$	0	0	0	0	1	10	0
$D_4–D_3^\#$	0	0	0	0	1	11	0
$D_4–D_3$	0	0	0	0	1	12	0

Note. Numerical values are provided for a *small* upward implicative interval, $C_4–D_4$ (2 semitones, major second), followed by realized intervals ranging in size from 12 semitones upward to 12 semitones downward. The higher the value, the stronger the expectancy (except for pitch proximity, where associations are predicted to be negative).

realized intervals ranging from 5 semitones upward (e.g., $D_4–G_4$) to 4 semitones downward (e.g., $D_4–A_4^\#$). All other realized intervals are unexpected (see Table 1, column 2, and Figure 2a–2d). For a large, upward implicative interval of 9 semitones (e.g., $C_4–A_4$), smaller realized intervals ranging from 5 semitones upward (e.g., $A_4–D_5$) to 6 semitones downward (e.g., A_4-$D_4^\#$) are expected; all others (e.g., $A_4–D_5^\#$, $A_4–C_4$) are unexpected (see Table 2, column 2, and Figure 2e–2h).

The third principle, REGISTRAL RETURN, describes a melodic archetype of the form X-Y-X or X-Y-X'. These three-tone archetypes exhibit "mirror" or "reflection" symmetry (or quasi-symmetry) in pitch about a point in time (i.e., the middle tone). The implicative and realized intervals are identical in these instances (e.g., $C_4–A_4–C_4$; 9 semitones up followed by 9 semitones down) or similar in size (e.g., $C_4–A_4–D_4$; 9 semitones up followed by 7 semitones down) but with a reversal in direction (upward to downward or vice versa). Because of the change in direction, similarity in size is defined as a difference of 2 semitones or less. Narmour believes that *exact* returns (complete symmetry) are more archetypal than *near* returns (quasi-symmetry), so the principle is graded accordingly. Realized intervals that are exact returns are quantified as 3. Near returns are quantified as 2 or 1. All other realized intervals have a value of 0 (see Tables 1 and 2, column 3, and Figure 2). Although REGISTRAL RETURN has been coded as a dichotomy in some tests of the I-R model (Cuddy & Lunney, 1995; Krumhansl, 1995; Krumhansl, Louhivuori, Toiviainen, Järvinen, & Eerola, 1999; Krumhansl et al., 2000; Thompson, Cuddy, &

Plaus, 1997; Thompson & Stainton, 1998), Narmour intended the principle to be graded (E. Narmour, personal communication, June 1991).

The fourth principle is called PROXIMITY. Though not articulated as clearly as Narmour's other principles, the theorist's intentions are relatively straightforward. The principle proposes that listeners have a general expectancy that realized intervals will be small, with expectancies increasing in strength as the realized interval becomes smaller and smaller. Because small intervals are defined as 5 semitones or smaller, the principle is coded as 6, 5, 4, 3, 2, 1, or 0 for realized intervals of 0, 1, 2, 3, 4, 5, or 6 or more semitones, respectively (see Tables 1 and 2, column 4, and Figure 2). This principle describes expectancies for proximity to the last tone listeners have heard with no consideration of the next-to-last tone. In other words, the principle describes determinants of expectancies that are instantiated at a simpler level than those described by the other principles in the I-R model, which consider the last *two* tones a listener has heard.

The final principle describes factors that contribute to a sense of finality in music. Called CLOSURE, this principle actually describes two separate factors that specify how two successive intervals contribute to the perception of melodic closure (as noted earlier): (a) a change in direction and (b) a reduction in interval size (using the rules described above for INTERVALLIC DIFFERENCE). Because the two factors are independent, they can occur jointly or on their own. The principle has a numerical value of 2 when both factors occur simultaneously, 1 when only one is operative, and 0 otherwise (see

Table 2
Quantification of the Principles From the Implication-Realization (I-R) and Two-Factor Models

Realized interval	I-R model					Two-factor model	
	Registral direction	Intervallic difference	Registral return	Proximity	Closure	Pitch proximity	Pitch reversal
A_4–A_5	0	0	0	0	0	12	−1
A_4–$G_5^{\#}$	0	0	0	0	0	11	−1
A_4–G_5	0	0	0	0	0	10	−1
A_4–$F_5^{\#}$	0	0	0	0	0	9	−1
A_4–F_5	0	0	0	0	0	8	−1
A_4–E_5	0	0	0	0	0	7	−1
A_4–$D_5^{\#}$	0	0	0	0	0	6	−1
A4–D_5	0	1	0	1	1	5	−1
A_4–$C_5^{\#}$	0	1	0	2	1	4	−1
A4–C_5	0	1	0	3	1	3	−1
A_4–B_4	0	1	0	4	1	2	−1
A_4–$A_4^{\#}$	0	1	0	5	1	1	−1
A_4–A_4	1	1	0	6	2	0	1
A_4–$G_4^{\#}$	1	1	0	5	2	1	1
A_4–G_4	1	1	0	4	2	2	1
A_4–$F_4^{\#}$	1	1	0	3	2	3	1
A_4–F_4	1	1	0	2	2	4	1
A_4–E_4	1	1	0	1	2	5	1
A_4–$D_4^{\#}$	1	1	0	0	2	6	1
A_4–D_4	1	0	1	0	1	7	2.5
A_4–$C_4^{\#}$	1	0	2	0	1	8	2.5
A_4–C_4	1	0	3	0	1	9	2.5
A_4–B_3	1	0	2	0	1	10	2.5
A_4–$A_3^{\#}$	1	0	1	0	1	11	2.5
A_4–A_3	1	0	0	0	1	12	1

Note. Numerical values are provided for a *large* upward implicative interval, C_4–A_4 (9 semitones, major sixth), followed by realized intervals ranging in size from 12 semitones upward to 12 semitones downward. The higher the value, the stronger the expectancy (except for pitch proximity, where associations are predicted to be negative).

Tables 1 and 2, column 5, and Figure 2). Incorporation of the principle into the model assumes that, all other things being equal, listeners expect closure or finality more than they expect openness or continued implication.

In summary, the I-R model provides detailed and quantifiable specification of five principles of melodic expectancy that are claimed to be innate. Listeners' expectancies are determined by the degree to which the hypothetical next tone in a melody adheres to each of the five principles. Because each principle can vary on its own, tones become more and more probable (and, hence, more expected) as they adhere to a larger number of principles.

On the one hand, the hardwired principles from the model are rooted in gestalt grouping laws, which implies that they are domain general. On the other hand, the model's principles are specified with such precision that extending them to other areas of audition, or to other modalities, is virtually impossible. For example, the arbitrary threshold between small and large intervals (relevant to four of the five principles) is difficult to relate to other domains. Moreover, the modularity perspective (Fodor, 1983) adopted by Narmour (1990)—who described the principles as *innate, hardwired, bottom-up, brute, automatic, subconscious, panstylistic,* and *resistant to learning*—implies that the principles are domain specific and independent of age and exposure to music.

Although the I-R model's claims about innateness are provocative, previous tests of the model's hardwired principles reported convergent and supportive findings. In all cases, the model proved to be statistically significant in multivariate analyses. Outcome

measures included ratings of how well individual test tones continued two-tone stimulus intervals (Cuddy & Lunney, 1995; Krumhansl, 1995) or actual melodies (Krumhansl, 1997; Krumhansl et al., 1999, 2000; Schellenberg, 1996) and production tasks that required participants to sing (Carlsen, 1981; Unyk & Carlsen, 1987) or to perform (Thompson et al., 1997) continuations to two-tone stimuli. Participants were adults from different cultural backgrounds (American, Canadian, Chinese, Finnish, German, Hungarian, and Sami) with varying amounts of musical training. Nonetheless, the I-R model's "success" in these cases was based on the null hypothesis of *no association.* In other words, the model performed better than one would expect if all of the variables were created with a random-number generator. It is clear that more stringent tests, or comparisons with alternative models, are required (see Cutting et al., 1992).

The model also contains much overlap (i.e., collinear principles), with different principles making similar predictions. For example, three principles (INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE) describe expectancies for small realized intervals. In other instances, the model's principles make contradictory predictions (e.g., for small implicative intervals, REGISTRAL DIRECTION and REGISTRAL RETURN are negatively correlated). These observations imply that the model is needlessly complex and overspecified. Indeed, in most tests of the model (Cuddy & Lunney, 1995; Krumhansl, 1995, 1997; Krumhansl et al., 1999, 2000; Schellenberg, 1996), at least one of its five predictors failed to make a significant contribution in multiple regression analyses.
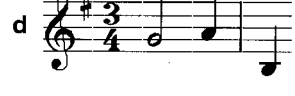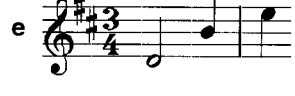
Figure 2. Combinations of possible implicative and realized intervals in musical notation. The first two tones constitute the implicative interval. The interval is "open" and implicative because, compared with the second tone, the first tone is longer in duration and more stable in the key (G major, a–d; D major, e–h), and the first tone occurs at a stronger metrical position in the measure. The second and third tones constitute the realized interval. Examples a–d illustrate a small implicative interval (2 semitones) followed by a small realized interval (a and b; 3 semitones) or a large realized interval (c and d; 10 semitones), which either maintains direction (a and c; upward/upward) or changes direction (b and d; upward/downward). Examples e–h illustrate a large implicative interval (9 semitones) followed by a small realized interval (e and f; 5 semitones) or a large realized interval (g and h; 8 semitones), which either maintains direction (e and g; upward/upward) or changes direction (f and h; upward/downward). The quantified expectancy values of the third tone (given the first two tones) are provided separately for each principle from both models. Higher values indicate stronger expectancies for all principles except PITCH PROXIMITY.

Presumably, general perceptual principles governing melodic expectancies would be relatively few in number. In line with this view, Schellenberg's (1997) simplified *two-factor* model reduced the core set of principles from five to two, which are called PITCH PROXIMITY and PITCH REVERSAL. Because the two factors are completely orthogonal (derived initially through principal-components analysis; see Schellenberg, 1997, for a detailed description), the two-factor model contains no redundancy. More importantly, the simplified model does not appear to sacrifice any of the predictive accuracy of the original model regardless of the particular experimental task used to test melodic expectancies. Specifically, Schellenberg's (1997) reanalyses demonstrated that the two-factor model equaled or outperformed the

I-R model at predicting responses across a variety of tasks and groups of participants, including (a) musically trained or untrained listeners who rated how well test tones continued tonal or atonal stimulus melodies (Schellenberg, 1996), (b) Chinese or American listeners who made continuation ratings for Chinese melodies (Schellenberg, 1996), and (c) musically trained or untrained listeners who rated how well test tones continued two-tone stimulus intervals (Cuddy & Lunney, 1995). Schellenberg (1996) also showed that the I-R model can be simplified without loss of predictive accuracy in explaining response patterns of music students from Germany, Hungary, and the United States who sang continuations to two-tone stimuli (Carlsen, 1981; Unyk & Carlsen, 1987).

In the spirit of the original I-R model, both principles of the two-factor model are rooted in the gestalt principle of proximity. As noted, pitch proximity is an important grouping factor in audition (Bregman, 1990). Tones that are proximate in pitch are grouped together. Conversely, tones far apart in pitch are unlikely to be grouped. For sequential tones (melodies), a predisposition for pitch-based streaming means that a tone with a pitch far removed from others in a sequence will often be perceived as coming from a different stream or source. Thus, any melody with large pitch distances between tones is relatively difficult to perceive as a unified gestalt.

The two-factor model's principle of PITCH PROXIMITY states simply that listeners expect subsequent tones in a melody to be proximate in pitch to tones they have already heard. Unlike the proximity principle of the I-R model, however, PITCH PROXIMITY assumes no arbitrary threshold between proximate and nonproximate tones. Rather, tones are said to become less and less expected as they move further away in pitch. Specifically, after any implicative interval in a melody, listeners expect the realized interval to be as small as possible, such that a unison (0 semitones, or a repetition of the second tone of the implicative interval) is the most expected interval, followed by realized intervals of ever increasing size (1 semitone, 2 semitones, etc.). The principle is quantified according to the size of the realized interval, in semitones for Western music (see Tables 1 and 2, column 6, and Figure 2), although any other logarithmic transformation of frequency into pitch (e.g., for non-Western cultures with nonsemitone scales) would work equally well. The principle assumes simply that melodies are perceived as groups of tones and that proximity is a fundamental grouping principle, as it is with other auditory signals and with visual stimuli. Because the principle has higher values for less proximate intervals, it should be negatively correlated with measures of expectancy. Similar ways of requantifying proximity have been adopted by other researchers (Krumhansl, 1995; Krumhansl et al., 1999, 2000).

The second principle of the two-factor model, called PITCH REVERSAL, describes expectancies that a melody will change direction (upward to downward/lateral or downward to upward/lateral). The principle incorporates aspects of REGISTRAL RETURN and REGISTRAL DIRECTION from the I-R model, as well as the *gap-fill* melodic process described originally by Meyer (1973; see also von Hippel, 2000) and verified experimentally by Schmuckler (1989). It is a "second-order" proximity-based principle, meaning that grouping principles based on proximity are instantiated at a relatively complex level. Whereas expectancies based on PITCH PROXIMITY are based solely on the last tone listeners have heard, PITCH REVERSAL considers the last two tones. Accordingly, PITCH REVERSAL requires more detailed processing and places greater demands on working and sensory memory.

PITCH REVERSAL describes two tendencies that contribute to expectancies for reversals. Both are modified versions of principles from the I-R model. One tendency describes particular melodic contexts in which reversals are expected. Listeners are said to expect that a melody will reverse direction after they hear two tones separated by a "large" implicative interval, retaining Narmour's definition of large (≥ 7 semitones). This tendency is identical to REGISTRAL DIRECTION except that it makes no predictions about pitch direction after *small* implicative intervals. Large intervals violate the basic expectancy for proximity and disrupt melodic

grouping. When direction is reversed immediately, melodic coherence is more likely to be restored. Realized intervals that reverse direction after a large implicative interval are quantified as 1 (e.g., Figure 2f); realized intervals that maintain direction have a value of –1 (e.g., Figure 2e; the 1 vs. –1 coding restricts this tendency to large intervals).

The second tendency describes expectancies for pitch reversals that produce patterns with mirror symmetry (or near symmetry) in pitch about a point in time. Specifically, listeners often expect that the next tone in a melody will be proximate in pitch ($\pm$ 2 semitones) to the first tone of an implicative interval, such that a symmetric structure (X-Y-X) or near-symmetric structure (X-Y-X') is formed. This tendency is identical to REGISTRAL RETURN except that it makes no distinction between exact and near returns. Because this type of symmetry occurs when two tones are proximate in pitch but separated by an intervening tone, it can be thought of as a higher order determinant of expectancies based on proximity. Tones proximate to the next-to-last tone are coded 1.5 (e.g., Figure 2b); others have a value of 0 (e.g., Figure 2a). This second tendency is coded with 1.5 and 0, rather than 1 and 0, so that it is weighted appropriately relative to the first tendency. When one considers all possible combinations of these two tendencies, PITCH REVERSAL can be quantified as –1, 0, 1, 1.5, or 2.5 (see Tables 1 and 2, column 7, and Figure 2). This coding method means that PITCH REVERSAL is essentially orthogonal to PITCH PROXIMITY ($r \approx 0$) regardless of the particular set of stimuli being examined.

In sum, the two-factor model describes two orthogonal variables that are said to influence melodic expectancies. In contrast to the I-R model, the two-factor model is agnostic with respect to the issue of innateness. Rather, both factors are rooted in a gestalt principle that is known to generalize widely, extending to audition in general, as well as to vision. Because grouping on the basis of proximity appears to be a perceptual predisposition, its immediate or eventual extension to music is likely to be mandatory. Whether such extensions are innate or acquired is a relatively moot point that is virtually untestable.

Despite their mutual emphasis on general predispositions, the I-R and two-factor models acknowledge that contextual factors play an important role in determining melodic expectancies. Though not described explicitly by either model, these contextual factors also influence whether a subsequent tone is expected and considered compatible with a particular musical context. For example, when a musical key is well established, some of the variance in expectancies unaccounted for by the I-R and two-factor models can be explained by relatively high-level and culture-specific variables, such as the *tonal hierarchy* (Krumhansl, 1990, p. 30) or conceptually related variables (see Cuddy & Lunney, 1995; Thompson et al., 1997). The tonal hierarchy is an index of the stability of tones in Western major- or minor-key contexts. *Do* (the tonic) has the highest value; tones from outside the key have the lowest values. When no key or mode is established, or when the context is relatively impoverished (e.g., only a few tones have been heard), relatively low-level and culture-general indices of tonal compatibility are more relevant. These include measures of consonance, such as the frequency-ratio index devised by Schellenberg and Trehub (1994b).

Although the two-factor model has matched or exceeded the explanatory accuracy of the I-R model across a variety of exper-

imental contexts and groups of listeners, each of these successes was based on reanalyses of previously collected sets of data. The model's predictive accuracy has yet to be tested prospectively. Another potential problem is that the two-factor model was initially data derived using response patterns obtained by Schellenberg (1996). Hence, its ability to generalize to new stimulus materials and different methods remains unknown. Moreover, other researchers (Krumhansl et al., 1999, 2000; Thompson et al., 1997) have reported that the two-factor model fails to match the predictive accuracy of the original I-R model, claiming that attempts to simplify the model are "premature" (Krumhansl et al., 1999, p. 187). In short, the two-factor model appears to be a promising alternative to the I-R model, but further comparisons are required.

In an alternative attempt at improving the I-R model's explanatory accuracy, Krumhansl et al. (1999, 2000) modified two of the model's five principles and *added* another two. When this seven-variable model was tested, at least two of the predictor variables failed to make a unique contribution in explaining response patterns (Krumhansl et al., 1999, 2000). Thus, this extended model fails to rectify the overlap and overspecification of the original I-R model.

Comparing the Models

Following Cutting et al. (1992), we evaluated and compared the I-R and two-factor models on the basis of their simplicity, scope, and selectivity. Because the two-factor model has two parameters compared with the I-R model's five, the two-factor model is simpler. Thus, if the two-factor model matches or exceeds the I-R model in tests of scope and selectivity, it is the better model.

The scope of the models was tested by examining their ability to predict random data. As noted, a model's scope is too broad if it succeeds at predicting random data. We generated 20 vectors of random data ($Ns = 263$). Each datum was a quantitative value that corresponded to a particular implicative interval paired with a particular realized interval. The pairings included 25 realized intervals (ranging from 12 semitones in the same direction as the implicative interval to 12 semitones in the opposite direction) for each of 10 implicative intervals (1–5 semitones, 7–11 semitones). In addition, we considered implicative unisons (0 semitones) paired with 13 realized intervals (another unison plus intervals from 1 to 12 semitones). The I-R model makes identical predictions for upward and downward implicative intervals, as does the two-factor model.

Multiple regression was used to test whether either model could successfully predict any of the 20 random vectors. Both models explained only 1 of 20 vectors at a statistically significant level ($\alpha < .05$). In other words, neither model explained random data any better than one would expect by chance. A second, more powerful test examined the F values generated by both models for each of the 20 vectors. If the average or median F statistic were greater than 1, the scope would appear to be too broad. For both models, however, the obtained F value was greater than 1 in only 6 of the 20 tests, and F values did not differ in magnitude between models. In short, the scope of neither model is so broad that it successfully predicts random data.

The main body of the present report was dedicated to testing the selectivity of the models, or their ability to predict data that they

should predict. We conducted two experiments that tested and compared the I-R and two-factor models among participants who varied in age. In Experiment 1, adults, 11-year-olds, and 8-year-olds rated how well individual test tones continued fragments of melodies. Presumably, higher ratings would be given to test tones consistent with listeners' expectancies, as determined by the melodic fragments. In Experiment 2, musically sophisticated 11-, 8-, and 5-year-olds sang continuations to two-tone stimuli, assuming that their continuations would begin with tones consistent with their expectancies, as determined by the stimulus intervals.

All participants in both experiments were exposed to Western music, although older participants obviously had more exposure than their younger counterparts. Previous research provides unequivocal evidence that music perception and performance abilities are influenced by maturity and increased exposure to music. For example, before puberty, children's perception of tone patterns is relatively culture free (Andrews & Dowling, 1991; Dowling, 1990; Krumhansl & Keil, 1982; Lynch, Eilers, Oller, & Urbano, 1990; Schellenberg, 2001; Schellenberg & Trehub, 1999; Trainor & Trehub, 1994; Trehub, Schellenberg, & Kamenetsky, 1999), although formal training in music accelerates the enculturation process (Morrongiello, 1992). Music performance abilities also improve with age and continued practice (Davidson, 1985; Dowling, 1984; Hargreaves, 1986; Howe, Davidson, & Sloboda, 1998; Miller, 1987).

The experiments were not designed to test whether the I-R or two-factor models embody truly hardwired and innate principles of melodic expectancy. Rather, the goal was to provide a test of the relative *efficacy* and *generality* of the models by examining listeners who varied in age. The more general model should provide a better description of melodic expectancies across a wide range of age and musical abilities. We predicted that responses would become more systematic and better explained by both models as participants increased in age, maturity, and exposure to music. On the basis of results from analyses of preexisting data, we also predicted that the two-factor model would match or exceed the explanatory accuracy of the I-R model. The relatively complex nature of the I-R model and its high degree of collinearity precluded predictions about developmental differences among its five predictors. Moreover, Narmour (1990) made no such predictions. For the two-factor model, the first-order/second-order distinction between the two factors led to two hypotheses: (a) The first-order proximity principle (PITCH PROXIMITY) will exert a larger influence on melodic expectancies compared with its second-order counterpart (PITCH REVERSAL) across development, and (b) the second-order principle will require longer time and more exposure to music to manifest itself completely.

Experiment 1: Continuation Ratings

The purpose of the present experiment was twofold: (a) to determine whether earlier findings using continuation ratings would be replicable with a new set of stimulus melodies and (b) to examine whether musical expectancies—as measured by the I-R and two-factor models—change over development, and if so, how. The stimuli were taken from Acadian folk-song collections (see Figure 3). Acadians are French-speaking Canadians from the Maritime (east coast) provinces. This musical genre was selected because it is clearly tonal and familiar sounding yet it was unlikely

Figure 3. The melodic fragments used in Experiment 1. The fragments were from Acadian folk songs. Each ended in an upward implicative interval.

that the participants would have heard the actual songs. To limit the duration of the testing session, we chose melodies that ended in *upward* implicative intervals only. Upward intervals are considered to be more implicative than their downward counterparts (Narmour, 1990), although previous findings suggest that the I-R and two-factor models of melodic expectancy explain response patterns similarly well for fragments ending in upward or in downward intervals (Schellenberg, 1997).

Method

Participants. The sample included 14 adults, 14 older children, and 32 younger children. The participants were recruited without regard to musical training, but all had everyday exposure to Western music. The adults were undergraduate students registered in an introductory psychology course who received partial course credit for their participation. Most ($n = 11$) had 5 years of music lessons or less ($M = 2$ years, 3 months). Two adults had more than 5 years of lessons ($M = 8$ years, 6 months), and 1 adult had 28 years of lessons.

The older children were 10- and 11-year-olds ($M = 10$ years, 11 months; $SD = 8$ months; range = 9 years, 11 months to 12 years, 1 month). Half had never taken formal music lessons; the other half had on average 2 years, 4 months of lessons. The younger children were 7 and 8 years of age ($M = 8$ years, 3 months; $SD = 5$ months; range = 7 years, 5 months to 8 years, 11 months). The majority (20 of 32) had no music lessons. Eight of the younger children had 1 year of music lessons or less ($M = 8$ months); the other 4 children had between 2 and 3 years of lessons ($M = 2$ years, 4 months). Four additional children in the younger group were tested but subsequently excluded from the final sample because their responses showed little or no variance across trials (see *Procedure*).

Apparatus. Stimulus melodies were created initially as musical instrument digital interface (MIDI) files using sequencing software (Cubase)

installed on a Power Macintosh computer (7100/66AV). The same computer controlled stimulus presentation and response recording with customized programs created with Hypercard (for adults and older children) and PsyScope 1.1 (for younger children; Cohen, MacWhinney, Flatt, & Provost, 1993). The MIDI files were output through a MIDI interface (Mark of the Unicorn MIDI Express) to a Roland JV-90 multitimbral synthesizer set to an acoustic piano timbre. Stimuli were presented at a comfortable volume with lightweight personal stereo headphones (Sony CD550) in a sound-attenuating booth (Eckel Industries). A window in the booth allowed listeners to see the computer monitor. Listeners used a mouse connected to the computer to initiate trials and to record their responses.

Stimuli. The stimulus melodies were four fragments taken from Acadian folk-song collections (see Figure 3). Each fragment came from a different song. Fragments consisted of 14 or 15 tones and were unambiguously in a major or minor key in Western music. They started at the beginning but ended in the middle of a melodic phrase. Subtle differences in amplitude—as performed by a trained musician on a MIDI keyboard—clarified the meter of the melodies. Each melody had a duple meter (2 beats per measure; 2/4 or 6/8 time signature), with tempi selected to be the most natural sounding to the experimenters. Two of the fragments ended in a small upward interval of 2 or 3 semitones (Figure 3—Melody 1 or 2, respectively). The other two fragments ended in a large upward interval of 9 or 10 semitones (Figure 3—Melody 3 or 4, respectively). The fragments were chosen so that the final interval (i.e., between the last two tones) was open and maximally implicative according to Narmour (1990). Specifically, compared with the last tone of the fragment, the penultimate tone had a longer duration and a stronger metrical position (greater intensity), and it was more stable in the key of the fragment (according to conventional music theory and the tonal hierarchy).

A set of 15 test tones was generated as possible continuations for each fragment. Each tone had identical duration and temporal location, which

corresponded to the note that followed the fragment in the actual folk song. For each fragment, the set of test tones included all tones in the key of the fragment that fell within an octave (12 semitones) from the final tone. Seven test tones were higher than the final tone, 7 were lower, and 1 repeated the final tone.

Procedure. The procedure followed that of Schellenberg (1996). Participants were tested individually and received instructions both verbally and on the computer screen. Their task was to rate how well individual test tones *continued* the stimulus melodies. Listeners were told specifically that we were not interested in how well the test tones *completed* the melodies. Adults and older children made ratings on a scale from 1 (*extremely poor continuation*) to 7 (*extremely good continuation*). Younger children used a 5-point pictorial scale, with each point matched to a schematic drawing of a face that ranged from very sad (rating of 1, corresponding to a very bad continuation) to very happy (rating of 5, corresponding to a very good continuation). Listeners were urged to use the entire scale, reserving the endpoints for extreme cases.

The procedure was first demonstrated to participants on the keyboard. The experiment began with a practice melodic fragment that was drawn from the same folk-song collections as the test melodies. The fragment was presented three times without a test tone. The fourth and subsequent presentations of the fragment were followed by a test tone, which listeners rated by clicking the mouse on the appropriate number (or picture) of the rating scale. Listeners made ratings for eight different test tones for the practice fragment. The trials were self-paced. After the practice trials, listeners were informed that the eight test tones they had rated were a representative sample of test tones they would hear in the actual experiment, and they were again urged to use the complete scale.

For the adults and older children, the testing session consisted of four blocks of trials, one for each of the four fragments. Each block was identical to the practice session except that listeners rated 15 test tones rather than 8. The order of the four blocks was randomized separately for each listener, as were the 15 test tones in each block (60 ratings in total). Because of their relatively limited attention spans, the younger children made 30 rather than 60 ratings. They rated 15 test tones for both of two fragments, one ending in a small interval (2 or 3 semitones), the other in a large interval (9 or 10 semitones). Half of the listeners heard the small-interval block before the large-interval block; the blocks were presented in reverse order for the other listeners. Four listeners were assigned to each of eight cells (2 small intervals × 2 large intervals × 2 orders), such that each of the 60 test tones was rated by 16 of the 32 children. Children whose responses did not vary or varied only slightly (i.e., between two adjacent values on the 5-point scale) were excluded from the final sample. The entire testing procedure lasted approximately 30 min for older children and adults and approximately 20 min for younger children.

Results and Discussion

Before discussing the analyses in detail, we summarize the main findings: (a) With increases in age and exposure to music, melodic expectancies became more systematic and better explained by both models; (b) the two-factor model consistently matched or exceeded the explanatory accuracy of the I-R model; (c) listeners of all ages expected the next tone in a melody to be close (proximate) in pitch to the tone heard most recently; and (d) with increases in age and exposure to music, expectancies became influenced by additional properties of the melodies. Specifically, adult listeners expected tones to be proximate to the penultimate tone they had heard. They also expected a shift in the pitch direction of the melody after hearing two tones separated by a large leap (i.e., a large implicative interval).

Average ratings. The first set of analyses examined ratings that were averaged across listeners within each age group. The data are illustrated in Figure 4. These analyses ignored individual differences within groups, focusing instead on overall response patterns. The experimental unit was the individual test tone. When each age group was analyzed separately, the outcome variable had 60 ratings (one for each test tone) averaged over 14 adults, 14 older children, or 16 younger children. When the three groups were analyzed simultaneously, average responses from the younger children were converted from a 5- to a 7-point scale to be comparable with the other two groups (as in Figure 4), and the outcome measure had 180 average ratings (60 from each of the three age groups). Average ratings were significantly correlated between groups (rs = .775, .656, and .575, Ns = 60, ps < .001, for adults and older children, adults and younger children, and older and younger children, respectively). Nonetheless, in each case, a substantial portion of the variance (40%–70%) was nonoverlapping and indicative of age-related differences in responding.

Pairwise correlations among predictor variables are provided in Table 3 (in regular roman type) separately for the two models. As noted in previous research (Schellenberg, 1996, 1997), the I-R model contains a set of intercorrelated terms (INTERVALLIC DIFFERENCE, PROXIMITY, and CLOSURE), whereas the two-factor model was designed to have orthogonal predictors. Table 4 presents simple associations between predictors and ratings. For the groups combined, the average ratings were significantly correlated with each predictor from both models. The results were less consistent when the groups were analyzed separately, particularly for the child groups.

Two hierarchical multiple regression models were used to predict listeners' ratings and to test the explanatory accuracy of the I-R and two-factor models. On the first step, two variables were entered. Both were designed to control for extraneous variance unrelated to either model, which, in turn, made our tests of the models more powerful. One was a blocking variable (melody) that partialed out differences in the magnitude of ratings across the four stimulus melodies. Such differences (statistically significant for 3 individual adults, 3 older children, and 4 younger children but not for any of the averaged sets of data) were of no theoretical interest. The other variable (tonal hierarchy) accounted for differences in the perceived stability of the various test tones in the key of the melody. After a key is established, even children as young as 6 years of age judge *do* (the most stable tone in a key) to fit better with the key than *ti* (an unstable tone; Cuddy & Badertscher, 1987). The variable consisted of the quantified values provided by Krumhansl (1990, Table 2.1, p. 30). The tonic (*do*) had the highest value, followed by the other tones in the tonic triad (*sol* and *mi*), and, finally, by the other tones in the key (*fa*, *la*, *re*, and *ti*). This variable was positively associated with average ratings from each of the three groups (rs = .339, .291, and .424 for adults, older children, and younger children, respectively). Compared with the two consonance variables proposed by Krumhansl et al. (1999, 2000), tonal hierarchy provided a better fit to the data for each age group (for adults, older children, and younger children, respectively, tonal hierarchy: adjusted R^2s = .100, .069, and .166; consonance variables: adjusted R^2s = .092, .049, and .032).

When the three groups were analyzed together, a third variable (age, treated categorically) was also included in the first step to control for differences in the magnitude of ratings across groups. Such differences were significant, $F(2, 177) = 14.63$, $p < .001$, $\eta^2 = .142$, but of little theoretical interest. The younger children
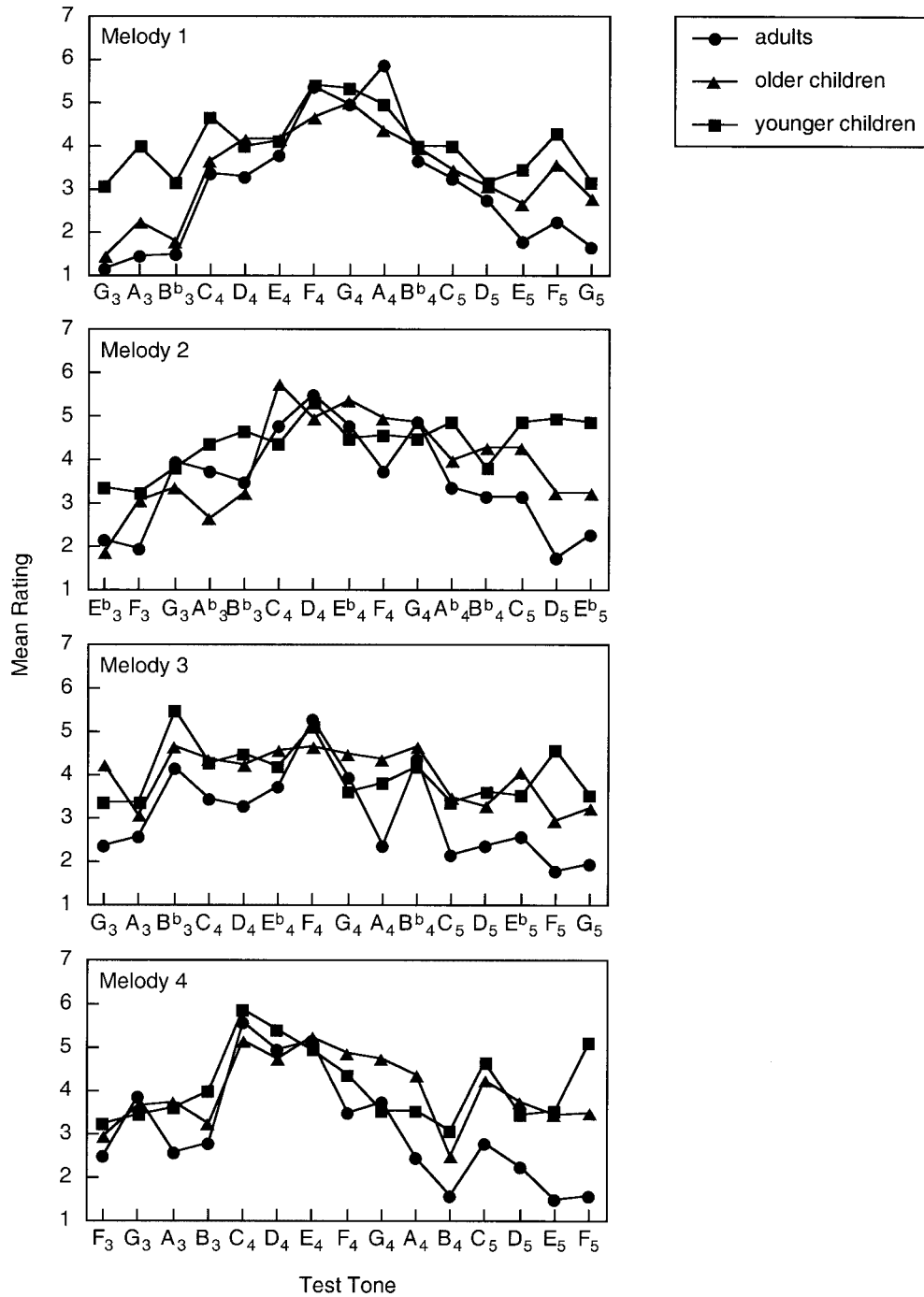
Figure 4.   Mean ratings for each of the 60 test tones in Experiment 1, shown separately for each melody and each age group.

gave the highest ratings on average ($M = 4.15$, when converted to a 7-point scale), followed by the older children ($M = 3.83$) and then the adults ($M = 3.19$). Ratings from younger children were significantly higher than those from adults ($p < .001$), as were ratings from older children ($p = .001$), but the two groups of children did not differ (Tukey's test). In short, compared with adults, children responded more favorably to all of the test tones.

The second step of the hierarchical regression provided the test of the two expectancy models. In one analysis, the I-R model's five predictors were entered; in another, the two predictors from the two-factor model were entered. Summary statistics are provided in Table 5. Both models were significant for each of the three age groups and the groups combined. In each case, the two-factor model resulted in no loss of explanatory power com-

Table 3
Pairwise Correlations Among Predictor Variables

Model and variable	Registral direction		Intervallic difference		Registral return		Proximity	
Implication-realization model								
Intervallic difference	.101	*143**						
Registral return	.020	*033*	.022	*036*				
Proximity	.000	*009*	.688*	*627**	.030	*−.027*		
Closure	.047	*012*	.401*	*323**	.120	*124**	.438*	*388**

	Pitch proximity	
Two-factor model		
Pitch reversal	−.073	*−.030*

Note. Values for Experiment 1 ($N = 60$) are in regular roman type. Values for Experiment 2 ($N = 525$) are in italics.
* $p < .05$.

pared with the I-R model. The variance explained was consistently higher for the two-factor model, despite the fact that this comparison was biased in favor of the I-R model (i.e., multiple R^2 always increases when additional variables are added to a regression model, even with random variables). Unbiased comparisons of the models (i.e., using adjusted R^2 values) revealed even greater advantages for the two-factor model (for adults, older children, younger children, and the groups combined, respectively, I-R model: adjusted R^2s = .645, .631, .287, and .579; two-factor model: adjusted R^2s = .768, .661, .324, and .625).

Finally, tests of Krumhansl et al.'s (1999, 2000) seven-variable model (not tabled) were also performed. These indicated that it failed to match the explanatory accuracy of the two-factor model for any age group (adjusted R^2s = .714, .642, and .105, for adults, older children, and younger children, respectively).[4] Their model predicted response patterns at better-than-chance levels for the adults and older children but not for the younger children.

Examination of the unique contributions of the individual regressors (see Table 5) proved to be consistent with previous indications that the I-R model is overspecified (Schellenberg, 1996, 1997). Although all of the principles except for CLOSURE were significant for the groups combined, analyses conducted separately for each age group revealed that INTERVALLIC DIFFERENCE and CLOSURE were not reliable for any group, and that REGISTRAL RETURN failed to make a unique contribution among either group of children. Indeed, for younger children, none of the I-R model's predictors was significant. For the two-factor model, both predictors made a significant unique contribution for the groups combined, and PITCH PROXIMITY was a robust predictor across age groups. PITCH REVERSAL was significant for adults but not for either group of children, a finding consistent with our claims about the principle's second-order status.[5]

The next set of analyses examined whether the predictor variables from the I-R model would interact with age. To test this possibility, we added five interaction terms individually to the regression model for the groups combined. Each term represented a two-way interaction between age and one of the predictors from the I-R model. Significant interactions were uncovered between age and PROXIMITY, $F(2, 166) = 11.26$, $p < .001$; age and INTER-

VALLIC DIFFERENCE, $F(2, 166) = 7.14$, $p = .001$; and age and CLOSURE, $F(2, 166) = 4.96$, $p = .008$. When these three two-way interaction terms were included simultaneously in the analysis, only the interaction between age and PROXIMITY remained significant, $F(2, 162) = 3.45$, $p = .034$. Inclusion of the single interaction term in the model increased the explained variance from 60.5% to 65.2%. The PROXIMITY predictor was stronger for adults than for younger children, $F(1, 108) = 20.30$, $p < .001$, and for older than for younger children, $F(1, 108) = 10.99$, $p = .001$. There was no difference between adults and older children (see Table 5). In short, the principle appears to reach adult levels of predictive accuracy some time between 8 and 11 years of age.

Two-way interactions with age were tested identically with the two-factor model. Inclusion of a two-way interaction term between age and PITCH PROXIMITY improved the fit of the model, $F(2, 169) = 16.66$, $p < .001$, as did the interaction between age and PITCH REVERSAL, $F(2, 169) = 4.78$, $p = .010$. Joint inclusion of the two interactions revealed that both made a unique contribution to explaining response patterns, $Fs(2, 167) = 16.23$ and 4.52, $p < .001$ and $p = .012$, respectively. Their inclusion in the model increased the variance explained from 64.2% to 71.6%. Follow-up pairwise comparisons revealed that PITCH PROXIMITY provided a more complete account of response patterns for adults compared with those for older children, $F(1, 110) = 4.39$, $p = .039$; for adults compared with younger children, $F(1, 110) = 31.18$, $p < .001$; and for older children compared with younger children, $F(1,$

---

[4] These analyses compared the fit of Krumhansl et al.'s (1999, 2000) extended model with that of the two-factor model plus tonal hierarchy. Whereas the two-factor model makes it clear that additional variables should be added according to the context, the extended model has two consonance variables that are said to apply across contexts.

[5] Additional analyses confirmed that ratings from the two melodies ending with a small implicative interval (Melody 1 and Melody 2) were predicted equally well by the two-factor model, as were ratings for both melodies ending with a large implicative interval (Melody 3 and Melody 4). Moreover, the predictive power of PITCH PROXIMITY and PITCH REVERSAL did not vary across the two stimulus melodies for either small or large implicative intervals.

Table 4
Simple Associations (Pearson Correlations) Between Predictor Variables and Average Ratings in Experiment 1

Model and variable	Registral direction	Intervallic difference	Registral return	Proximity	Closure
Implication-realization model					
Adults ($n = 60$)	.197	.578*	.289*	.678*	.409*
Older children ($n = 60$)	.191	.538*	.196	.689*	.293*
Younger children ($n = 60$)	.168	.344*	.140	.380*	.162
Groups combined ($N = 180$)	.170*	.458*	.200*	.549*	.280*

	Pitch proximity	Pitch reversal
Two-factor model		
Adults ($n = 60$)	−.746*	.390*
Older children ($n = 60$)	−.725*	.202
Younger children ($n = 60$)	−.399*	.206
Groups combined ($N = 180$)	−.591*	.256*

* $p < .05$.

110) $= 12.88$, $p < .001$ (see Table 5). PITCH REVERSAL was also a better predictor among adults compared with older children, $F(1, 110) = 6.60$, $p = .012$, and among adults compared with younger children, $F(1, 110) = 6.87$, $p = .010$, but it was equally weak for the two groups of children. Thus, the interaction between age and PITCH REVERSAL is consistent with our suggestion that this principle becomes operative relatively late in development (after age 11).

The interaction between age and PITCH PROXIMITY could reflect an even longer developmental trajectory—which we did not predict—as this principle became stronger with increasing age. Nonetheless, this increase could also stem from more systematic responding among older participants. For example, the ability to attend consistently to the task at hand would obviously increase with age. Moreover, compared with PITCH REVERSAL, PITCH PROXIMITY accounted for a larger unique proportion of the variance in the data for each age grouping (see Table 5), as predicted. In fact, PITCH PROXIMITY made a stronger contribution for our youngest

group of listeners (8-year-olds) than PITCH REVERSAL did for our oldest group (adults).

Individual listeners. The next set of analyses examined individual response patterns separately for each adult, older child, and younger child. These analyses ensured that the results from the averaged data were not an artifact of eliminating variance due to individual differences. Specifically, multiple regression was used to determine how well the I-R and two-factor models could explain response patterns from each individual listener (i.e., the 60 ratings from each adult or older child, or the 30 ratings from each younger child). For both models, the two variables from Step 1 of the previous analyses (i.e., melody and tonal hierarchy) were included to hold constant any effects they had on response patterns.

Summary statistics are provided in Table 6 separately for each age group. The predictive accuracy of the two models was compared across age groups by analyzing adjusted R^2 values obtained from individual listeners. A 2 × 3 mixed-design analysis of

Table 5
Summary of Hierarchical Regression Models Fit to Average Data in Experiment 1

Step and variable	Adults			Older children			Younger children			Groups combined		
	R^2	ΔR^2	sr^2	R^2	ΔR^2	sr^2	R^2	ΔR^2	sr^2	R^2	ΔR^2	sr^2
Step 1	.129	.129		.158*	.158*		.207*	.207*		.255*	.255*	
Melody			.014			.073			.027			.015
Tonal hierarchy			.107*			.080*			.166*			.091*
Age												.141*
Step 2: I-R model	.699*	.570*		.687*	.529*		.396*	.189*		.605*	.350*	
Registral direction			.031*			.030*			.026			.024*
Intervallic difference			.023			.005			.019			.012*
Registral return			.032*			.015			.001			.013*
Proximity			.093*			.210*			.023			.082*
Closure			.013			.014			.000			.000
Step 2: Two-factor model	.791*	.662*		.696*	.538*		.393*	.186*		.642*	.387*	
Pitch proximity			.510*			.504*			.144*			.324*
Pitch reversal			.113*			.018			.032			.043*

Note. sr^2 = the proportion of variance explained uniquely by a predictor variable. I-R = implication-realization.
* $p < .05$.

Table 6
Results From Regression Models Fit to Individual Listeners' Ratings in Experiment 1

Measure	Implication-realization model			Two-factor model		
	Adults	Older children	Younger children	Adults	Older children	Younger children
Mean R^2	.484	.398	.347	.494	.347	.245
Standard deviation R^2	.092	.134	.151	.129	.161	.147
Maximum R^2	.585	.588	.649	.658	.581	.595
Minimum R^2	.323	.197	.128	.187	.103	.003
Mean adjusted R^2	.391	.290	.163	.437	.273	.143

Note. Adults ($n = 14$) and older children ($n = 14$) made 60 ratings. Younger children ($n = 32$) made 30 ratings.

variance that included one repeated measure (model: I-R or two-factor) and one between-subjects variable (age: adults, older children, or younger children) uncovered a significant main effect of age, $F(2, 57) = 15.37$, $p < .001$, $\eta^2 = .350$. As shown in Table 6, ratings were more completely explained as listeners increased in age. The main effect of model was not significant, but there was a significant interaction between model and age, $F(2, 57) = 3.46$, $p = .038$, $\eta^2 = .108$. The two-factor model outperformed the I-R model for adults, $t(13) = 2.23$, $p = .044$, but the models did not differ for either child group.

The present investigation is the first to examine melodic expectancies among listeners of different ages. The findings extend those reported previously (Schellenberg, 1996, 1997). For all age groups, the two-factor model did not result in any loss of explanatory power compared with the I-R model. To illustrate, when the data were averaged across listeners, the two-factor model accounted for 63% of the variance compared with the I-R model's 59%. Even if the two-factor model were simply to match the I-R model in terms of its selectivity, its greater simplicity makes it the superior model of the present set of data. The two-factor model was also more sensitive to age-related changes in melodic expectancies as measured by the continuation-ratings task. PITCH PROXIMITY was progressively stronger with increases in age (i.e., stronger for adults than for both groups of children, stronger for 11- than for 8-year-olds), whereas PITCH REVERSAL was stronger among adults than among 11- and 8-year-olds. PROXIMITY as coded in the original I-R model was also sensitive to distinctions between younger children and the two older groups, but not between older children and adults. This finding provides additional evidence that the revised predictor variables are more sensitive to developmental changes than their original counterparts.

Experiment 2: Sung Continuations

In the present experiment, we used a production task to examine further melodic expectancies in childhood and to determine whether the findings from Experiment 1 would extend to a different method and a different set of stimuli. Instead of making continuation ratings as in the previous experiment, children were asked to sing continuations to melodic intervals (Adachi & Carlsen, 1995/1996). On each trial, children heard two tones and sang how they thought the "song" would continue, a method that has been used successfully with adults (Carlsen, 1981; Unyk & Carlsen, 1987). Although participants' continuations typically

comprised several tones, only the first tone was analyzed because expectancies for subsequent tones would be influenced by participants' intervening self-generated tones, as well as by the stimulus intervals. Compared with the ratings method of Experiment 1, the sung-continuation method is less constrained, allowing for freedom of expression in front of a captive audience (the experimenter), which would help to maintain young children's interest and attention. The method also guarantees that stimuli vary more from trial to trial compared with the ratings method.

Children from three age groups were tested: 11-, 8-, and 5-year-olds. The relatively difficult and musical nature of the task motivated us to prescreen children to ensure that their singing abilities would be commensurate with their intentions (Adachi & Carlsen, 1995/1996). In other words, we attempted to verify that the tones and intervals the children sang were those that they intended to sing. Thus, the sample was relatively sophisticated in their musical abilities and, consequently, not directly comparable with the children tested in Experiment 1. In fact, we expected the melodic expectancies of the present sample to be precocious in comparison.

Method

Participants. The final sample included 45 children, 15 from each of three age groups. The 11-year-olds were between 10 years, 10 months and 11 years, 11 months ($M = 11$ years, 6 months; $SD = 4$ months). The 8-year-olds were between 8 years, 0 months and 8 years, 9 months ($M = 8$ years, 5 months; $SD = 2$ months). All of the 11- and 8-year-olds had a history of music lessons ($M = 5$ years, 6 months and 2 years, 2 months, respectively), and all were currently taking lessons or participating in choirs. The 5-year-olds were between 4 years, 3 months and 6 years, 0 months ($M = 5$ years, 3 months; $SD = 7$ months). Six of the 5-year-olds had taken music lessons ($M = 1$ year, 2 months).

The sample was culled from a larger group of 91 children (twenty-two 11-year-olds, seventeen 8-year-olds, and fifty-two 5-year-olds). All had been preclassified as relatively "musical" compared with their peers by a music teacher or a musician. We examined their musical abilities in a screening procedure that required them to alternate with the experimenter in singing consecutive portions of familiar songs (see *Procedure*). Although 63 children successfully met the screening criterion, 18 were subsequently excluded from the test session because they failed to understand the experimental task (see *Procedure*). A smaller proportion of the 5-year-olds who were originally recruited were included in the final sample (29%) compared with the older age groups (11-year-olds: 68%; 8-year-olds: 88%), $\chi^2(2, N = 91) = 22.15$, $p < .001$. This imbalance should reduce age-related differences in expectancies and make tests of developmental trends more conservative than they would be otherwise.

Apparatus. A Sony Model CFS-W304 cassette tape player was used to present stimuli. Participants' continuations were recorded with a Sony F-98 cardioid microphone and a Marantz PMD 430 cassette tape recorder. Stimuli were digitally edited on a Macintosh IIvX computer with Sound Designer II software.

Stimulus materials. The stimuli comprised 25 different melodic intervals, each of which was presented at two different pitch levels (50 intervals in total). Each interval was an integer multiple of a semitone. The stimulus set included all intervals between 12 semitones ascending and 12 semitones descending. The second pitch of each interval fell between E_4 and $A_4^{\#}$, in the middle of young children's singing range (Drexler, 1938; Wassum, 1979). All intervals were sung by a female singer (Mayumi Adachi) without vibrato using the syllable /la/. Trained musicians verified that her pitch was accurate.

Stimulus intervals were recorded onto cassette tape. Two tapes were prepared. Each had 3 practice trials followed by 50 test trials (1 trial for each stimulus). Each trial began with an announcement of the trial number, followed by eight clicks of a metronome presented at one click per second. The first two clicks indicated the meter. The first and second tones of the sung stimulus interval were presented simultaneously with the third and fourth click, respectively. The remaining four clicks marked the beat for participants' sung continuations. After the eighth click, there was a 3-s pause before the next trial was announced. The 50 test trials consisted of two blocks. An initial block comprised all 25 different melodic intervals presented in random order (constrained so that direction changed from trial to trial), with pitch level selected randomly on each trial. A second block had another set of the 25 intervals (different random order) at the alternate pitch level. The two stimulus tapes had different random orders.

Procedure. The procedure was a modified version of the sung-continuation method, adapted by Adachi and Carlsen (1995/1996; for a detailed description, see Adachi, 1995) to make it appropriate for testing children. Children were tested individually in a quiet room. The entire session was recorded on cassette tape.

The session began with a screening procedure. Children were asked initially to identify renditions of "Twinkle Twinkle Little Star" followed by "Row Row Row Your Boat," which were sung with the syllable /la/. They were then asked to play a "music game" (illustrated in Figure 5, upper row) that required them to take turns with the experimenter while singing the same two songs using /la/. A visual guide had pink facelike tokens that represented notes the experimenter would sing and blue tokens that rep-

resented the child's notes. To proceed to the test session, children had to sing correctly the blue-token sections of "Twinkle Twinkle" and "Row Row" in two different musical keys.

Before children began the test session, we verified that they understood the experimental task (illustrated in Figure 5, lower row). The experimenter provided the following instructions:

> Here is a mystery song that you and I are going to sing. I will sing the beginning of this song, which is these pink ones (pointing to the pink tokens), and I would like you to continue this song by singing these blue ones (pointing to the blue tokens). As you can see, there are only four blue ones here, but you can sing as long as you want as if there were many more invisible blue ones (pointing to the right of the blue tokens).

Children were given four practice trials that included both small and large ascending and descending intervals. Children who began their continuation by *repeating* the stimulus interval (rather than *continuing* it) were given an additional two practice trials. Any child who repeated the interval more than twice during the practice trials was excluded from the sample.

Before beginning the test session, the children were informed that a tape would take the part of the experimenter. They were also told about the metronome clicks that marked the structure of the trials. Each child was assigned randomly to one of the two stimulus tapes. After three practice trials with the tape, the experimenter let the tape run through the 50 test trials. The procedure was identical for the three age groups except that the 5-year-olds took a short break after 25 trials. The entire session took approximately 30 min for 5-year-olds and between 15 and 30 min for older children.

Results and Discussion

We begin with a summary of the major findings. The results largely replicated those from Experiment 1. For all three age groups, the two-factor model exceeded the predictive accuracy of the I-R model. For each group, PITCH PROXIMITY explained a relatively large proportion of variance in responding, and its predictive strength did not change with age. As in Experiment 1, higher order grouping factors (as embodied in PITCH REVERSAL) increased in
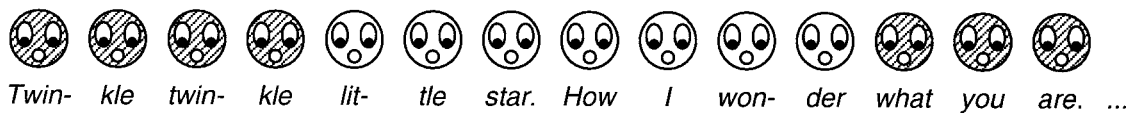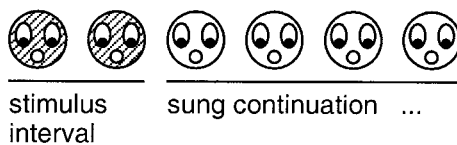

Figure 5. Illustrations used to explain the screening and test procedures to children in Experiment 2. In the screening procedure (upper row), the child and the experimenter took turns singing a familiar song using the syllable /la/. The hatched (pink in the display shown to the children) tokens represented the experimenter's part, whereas the clear (blue in the display) tokens represented the child's part. Only the first line of the song is illustrated, but the child and experimenter sang the whole song. In the test procedure (lower row), the child heard a stimulus interval (hatched tokens) and was asked to sing a continuation (clear tokens). The stimulus interval and the child's response were sung with the syllable /la/. Only the first tone of the continuations was analyzed.

importance later in development. In contrast to the previous experiment, however, the predictive accuracy of both models did not increase as the children became older. In fact, response patterns were explained equally well for 5-year-olds as for 11-year-olds. This result is likely to be a consequence of our inclusion criteria, which guaranteed that the youngest children were the most musically sophisticated for their age.

To prepare the data for statistical analyses, we assigned the first sung tone to 1 of 25 response-interval categories, separately for each continuation from each child. The response (realized) interval was the distance in pitch (in semitone multiples) between the second tone of the stimulus (implicative) interval and the first tone of the sung continuation. Categories ranged from 12 semitones downward to 12 semitones upward. Responses outside of this range were rare. The reliability of this process was confirmed by having an independent rater (a trained musician) categorize 14% of responses (seven from each participant). Cohen's kappa statistic was consistently high (κ = .95, .91, and .85 for 11-, 8-, and 5-year-olds, respectively). Twenty data points (< than 1%) were discarded (1, 8, and 11 for 11-, 8-, and 5-year-olds, respectively) for the following reasons: (a) The response interval was larger than an octave, (b) the pitch of the first sung tone was unclear, or (c) the participant failed to attend to the task.

Group data. In initial analyses, the response interval was regressed onto the stimulus interval separately for each age group. The data are illustrated in Figure 6. The figure confirms that a variety of response intervals was sung for each stimulus interval and that all 25 "possible" responses were indeed produced by the children. The figure also illustrates that many responses began with tones proximate to the second stimulus tone, as evidenced by the concentration of data along the horizontal midpoint in each figure. This pattern is consistent with both the I-R and two-factor models, which propose that listeners expect the next tone in a melody to be proximate in pitch to the tone heard most recently. In addition, the data in the upper-left and lower-right quadrants indicate that reversals of direction were more common when the children were asked to sing continuations to large rather than to small stimulus intervals. Finally, the data along the upper-left/lower-right diagonal indicate that singers began many of their pitch reversals with tones proximate to the first stimulus tone. These observations are in line with predictions from the I-R model (REGISTRAL DIRECTION and REGISTRAL RETURN) and with those from the two-factor model (PITCH REVERSAL). The correlation between the stimulus (implicative) and response (realized) intervals was significant for each group (rs = –.612, –.662, and –.449; Ns = 749, 742, and 739, ps < .001, for 11-, 8-, and 5-year-olds, respectively). The negative associations reflect the fact that upward (positive) responses were more likely to follow large, downward stimulus intervals, whereas downward (negative) responses were more likely after large, upward intervals.

Figure 6 also makes it clear that simply regressing the response interval on the stimulus interval precludes tests of the influence of proximity, which is central to both models. Proximate responses are represented by points clustered around the horizontal axis, which is the null hypothesis in regression. As an alternative, we treated the stimulus and response intervals categorically, following previous analyses of melodic expectancies measured with production tasks (Schellenberg, 1996; Thompson et al., 1997). Specifically, for each age group, a two-way data matrix was formed
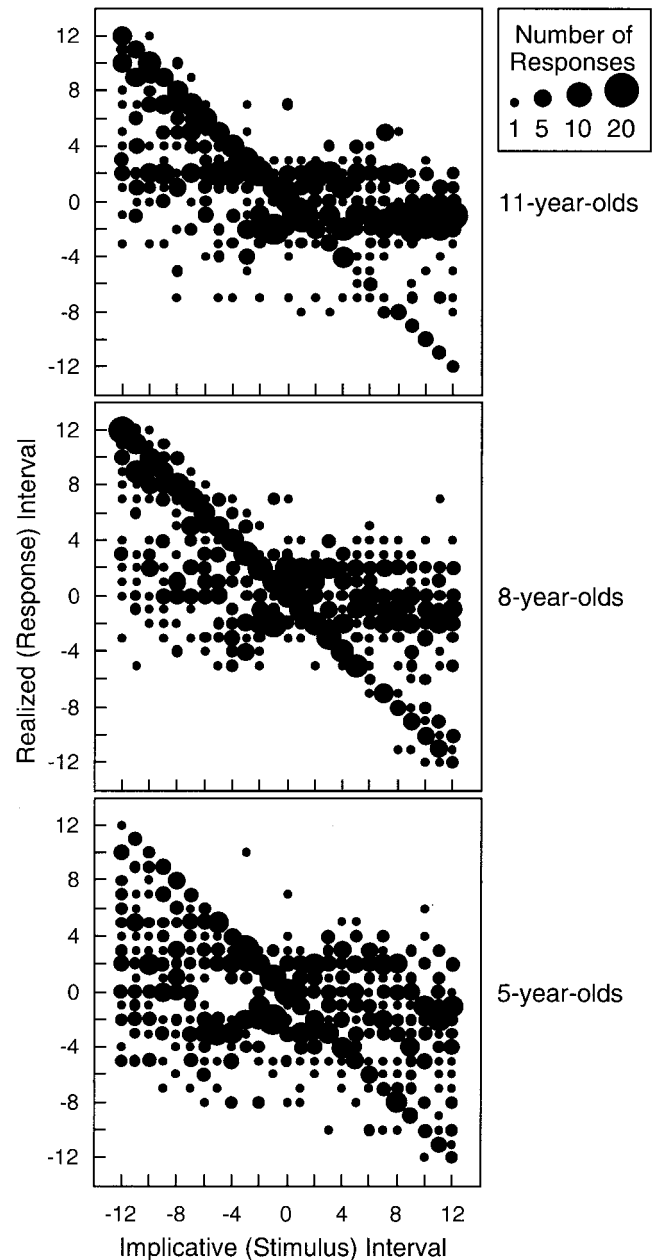
Figure 6. Realized (response) intervals plotted separately for each age group in Experiment 2 as a function of the implicative (stimulus) interval. The size of the circles indicates the number of responses (larger circles = more responses).

consisting of 25 stimulus intervals and 25 response intervals (625 cells). Each cell in the matrix contained a *frequency* value (corresponding to the area of the circles in Figure 6), representing the number of times a particular response interval was sung for a particular stimulus interval. The 2,230 responses (25 stimulus intervals × 2 responses per child × 3 groups × 15 children per group, less 20 missing values) were distributed among the 625 cells, constrained by the design so that each of the 25 stimulus intervals had 30 responses from each age group. Because Narmour

(1990) considered tritones (6 semitones) to be neither small nor large, and octaves (12 semitones) to be unique because of octave equivalence, these stimulus intervals (in upward and downward versions) were excluded in tests of the models. Hence, subsequent analyses included 525 cells for each age group. Rank-order (Spearman) correlations among age groups were similar in magnitude to the associations among groups observed in Experiment 1 (r_ss = .657, .644, and .668, Ns = 525, ps < .001, for 11- and 8-year-olds, 11- and 5-year-olds, and 8- and 5-year-olds, respectively).

The data were analyzed with multinomial log-linear models (following Schellenberg, 1996; Thompson et al., 1997), which are similar to multiple regression models but appropriate for categorical outcome variables and frequency data (Agresti, 1990). Models and individual predictors were tested by examining *deviance* (i.e., G^2, the likelihood ratio statistic, chi-square distribution), which is a measure of unexplained variance (i.e., the difference between observed and predicted frequencies). Specifically, the resulting decrease in deviance from baseline (equal cell frequencies) after adding the model's predictors was tested for significance. We then tested individual predictors by removing them from the model one at a time and noting whether the resulting increase in deviance was significant. Because of the large degree of power afforded by the large sample size (defined by the number of cells in the stimulus–response matrix), we adopted a relatively conservative alpha level (.01 instead of .05). In the present data set, with α = .05, a predictor variable could make a "statistically significant" contribution to a model yet account for less than 1/1,000th of the variance in response patterns.

Pairwise correlations among predictor variables are provided in Table 3 (in italics) separately for the I-R and two-factor models. The increase in sample size (from 60 to 525) meant that the predictors from the I-R model showed even more collinearity than they did in Experiment 1. The two predictors from the two-factor model remained orthogonal. Simple associations between each of the predictors and cell frequencies are provided in Table 7. The statistical power associated with a large sample size meant that all of these associations were significant and in the direction (positive or negative) specified by the models.

Before we tested and compared the expectancy models, we examined variables that were likely to explain extraneous variance in response patterns unrelated to either the I-R or the two-factor model. The tonal hierarchy variable from the previous experiment accounted for a significant proportion of the variance for each of the age groups (10.4%, 13.8%, and 9.9% for 11-, 8-, and 5-year-olds, respectively), based on the assumption that children perceived the underlying key to be the major key of the first stimulus tone. This assumption was problematic, however, particularly for instances in which a minor-key interpretation was more likely (e.g., when the stimulus interval was a minor third, or 3 semitones, upward).

As an alternative, we tested a variable that measured the degree of consonance between the first stimulus tone and the first tone of participants' sung responses. Specifically, we used Schellenberg and Trehub's (1994b) index for quantifying the frequency ratio of musical intervals. The index yields higher values for intervals with small- rather than large-integer frequency ratios. Consonant intervals such as octaves (12 semitones, frequency ratio 2:1), perfect fifths (7 semitones, 3:2), and perfect fourths (5 semitones, 4:3) have small-integer ratios and relatively high values on the index

(.910, .621, and .514, respectively), whereas dissonant intervals such as major sevenths (11 semitones, 15:8) and tritones (6 semitones, 45:32) have large-integer ratios and low values (.319 and .230, respectively). Infants, children, and adults exhibit advantages for consonant intervals compared with dissonant intervals, with the former being better perceived and remembered (Schellenberg & Trehub, 1994a, 1994b, 1996a, 1996b; Trainor, 1997). Infants also prefer to listen to consonant rather than to dissonant combinations of tones (Trainor & Heinmiller, 1998; Zentner & Kagan, 1996), and they perceive two different intervals to be similar if both are consonant (Schellenberg & Trainor, 1996). Although this index was correlated with tonal hierarchy (r = .778, N = 525, p < .001), as it is in general (Schellenberg & Trehub, 1994b), it was considered superior because it did not require us to assume that children always perceived the underlying key to be major, and it yields similar values when responses conform to a major or a minor key. In short, the ratio index variable provided an index of compatibility between responses and the stimulus context that was culture free and scale free, and conceptually unrelated to either model. Moreover, compared with tonal hierarchy, ratio index accounted for a larger proportion of variance in response patterns for each of the three age groups (16.7%, 22.6%, and 15.6% for 11-, 8-, and 5-year-olds, respectively),[6] thereby increasing power for our tests of the models.

As in Experiment 1, the principal analyses were hierarchical. Summary statistics are provided in Table 8. The baseline model (i.e., before the first step) included a variable that accounted for equal frequencies for each stimulus interval (30 for each age group). For analyses of the combined groups, an additional variable accounted for equal frequencies across age groups. On the first step of the hierarchical analysis, the ratio index was entered. The variable was positively and significantly associated with cell frequencies for each age group and for the groups combined. On the second step, we tested the explanatory accuracy of the I-R model in one analysis and the accuracy of the two-factor model in another.

The I-R model explained response patterns at above-chance levels for each age group and for the groups combined. To provide measures of effect size that are comparable with those from Experiment 1, we calculated proportion of deviance explained, which we refer to as R^2* for models (analogous to R^2), and sr^2* for predictors (analogous to sr^2), following previous reports (Schel-

---

[6] Additional analyses examined the possibility that the perceived key was the minor key of the first stimulus tone or the major or minor key of the second stimulus tone. In each case, tonal hierarchy accounted for a smaller proportion of the variance in response patterns compared with our initial approach (i.e., perceived key corresponds most closely to the major key of first stimulus tone). We also examined whether the frequency ratio between the second stimulus tone and the first response tone helped to explain response patterns. This additional variable did not substantially improve predictive accuracy across age groups, and it was not better in any case than the ratio index. Finally, we tested the two consonance variables used by Krumhansl et al. (1999, 2000); this pair of variables explained slightly more variance than did the ratio index for the 11-year-olds (19.6%), but much less for the 8- and 5-year-olds (12.5% and 10.3%, respectively). In line with our efforts to describe melodic expectancies as parsimoniously but as completely as possible, we elected to use the ratio index in the first step of our hierarchical analyses.

Table 7
*Simple Associations Between Predictor Variables and Cell Frequencies in Experiment 2
(all ps < .01)*

Model and group	Registral direction	Intervallic difference	Registral return	Proximity	Closure
Implication-realization model					
11-year-olds	.164	.401	.324	.518	.269
8-year-olds	.174	.380	.423	.486	.255
5-year-olds	.141	.487	.289	.527	.266
Groups combined	.160	.420	.347	.510	.263

	Pitch proximity	Pitch reversal
Two-factor model		
11-year-olds	−.537	.331
8-year-olds	−.489	.404
5-year-olds	−.582	.275
Groups combined	−.534	.337

Note. Values are the square root of the proportion of deviance (G^2) explained and therefore analogous to the correlations (Pearson's *r*) presented in Table 4 for Experiment 1.

lenberg, 1996; Thompson et al., 1997). All of the predictors except for CLOSURE made a unique contribution in each analysis, although some of these effects were small in magnitude (i.e., except for PROXIMITY, the unique explanatory contribution of each predictor was 6% or less).

The two-factor model proved to be highly significant for each age group and for the groups combined, as were both of its predictor variables. More importantly, the two-factor model did not sacrifice explanatory accuracy compared with the I-R model. Although the tests were biased in favor of the I-R model (i.e., additional predictors can only increase a model's fit to any set of data), the simplified model slightly outperformed the original model for each of the three age groups and for the groups combined. Moreover, individual predictors made much larger unique contributions to the two-factor model than did the predictors from the I-R model. Tests of the extended model proposed by Krum-

hansl et al. (1999, 2000) revealed its explanatory power to be relatively weak across age groups, accounting for 42.3%, 35.2%, and 44.4% of the deviance for 11-, 8-, and 5-year-olds, respectively. In fact, it failed to match the predictive accuracy of the original I-R model (plus ratio index) in any of the analyses.

Tests of interactions between age group and the predictors were conducted on the combined data set. For the I-R model, each possible two-way interaction term was added individually to the main-effects model (as in Experiment 1). Only one of the five interactions—between age and REGISTRAL RETURN—was reliable, $G^2(2, N = 1,575) = 11.8, p < .001$. Follow-up pairwise comparisons of the age groups revealed that 11-year-olds did not differ significantly from either of the younger groups, although REGISTRAL RETURN was a better predictor of the 8-year-olds' responses than of 5-year-olds' ($p < .001$; see Table 8). In short, age-related differences in the strength of the predictor were uninterpretable.

Table 8
Summary of Log-Linear Models Fit to Frequency Data in Experiment 2

Step and variable	11-year-olds			8-year-olds			5-year-olds			Groups combined		
	R^{2*}	ΔR^{2*}	sr^{2*}	R^{2*}	ΔR^{2*}	sr^{2*}	R^{2*}	ΔR^{2*}	sr^{2*}	R^{2*}	ΔR^{2*}	sr^{2*}
Step 1	.167*	.167*		.226*	.226*		.156*	.156*		.182*	.182*	
Ratio index			.167*			.226*			.156*			.182*
Step 2: I-R model	.505*	.338*		.562*	.336*		.500*	.344*		.515*	.333*	
Registral direction			.047*			.060*			.028*			.045*
Intervallic difference			.005*			.004*			.029*			.010*
Registral return			.028*			.056*			.014*			.031*
Proximity			.126*			.120*			.086*			.111*
Closure			.001			<.001			.001			<.001
Step 2: Two-factor model	.528*	.361*		.567*	.340*		.534*	.378*		.537*	.354*	
Pitch proximity			.315*			.271*			.351*			.309*
Pitch reversal			.077*			.107*			.045*			.075*

Note. R^{2*} = the proportion of deviance explained by a model; sr^{2*} = the proportion of deviance explained uniquely by a predictor variable; I-R = implication-realization.
* $p < .01$.

For the two-factor model, the interaction between age and PITCH REVERSAL was reliable, $G^2(2, N = 1,575) = 20.6$, $p < .001$. PITCH REVERSAL provided a better explanation of sung responses of 11-year-olds than it did for 5-year-olds, and for 8-year-olds compared with 5-year-olds, G^2s$(2, Ns = 1,050) = 8.2$ and 19.7, respectively, $ps < .005$. The principle did not differ in explanatory strength between the two older groups. The lack of an interaction between age and PITCH PROXIMITY indicates that the proximity predictor was similarly strong among all age groups. By contrast, PITCH REVERSAL appeared to have an extended developmental trajectory. Even among children who were selected to be musically advanced, the principle was not fully developed until somewhere between 5 and 8 years of age. This extended trajectory for PITCH REVERSAL is consistent with our hypothesis and with the interaction between age and PITCH REVERSAL reported in Experiment 1. The fact that the predictive accuracy of the principle plateaued at an earlier age (before 8 years of age) compared with the previous experiment (after 11 years) is likely to be due to the musical sophistication of the present sample of children, although other differences between experiments (i.e., stimuli, methods, covariates) might also have played a role.

Individual singers. Multivariate analyses of each child's data were inappropriate (42 responses in 525 cells). As an alternative, we conducted univariate analyses separately for each child for each predictor variable (see Table 9) to verify that patterns observed in the group data were similar to those observed across individuals. Specifically, we dichotomized each predictor variable from the I-R model (REGISTRAL DIRECTION and INTERVALLIC DIFFERENCE: 1 vs. 0; REGISTRAL RETURN: 1–3 vs. 0; PROXIMITY: 1–6 vs. 0; CLOSURE: 1 or 2 vs. 0) and the two-factor model (PITCH PROXIMITY: 0–5 vs. 6–12; PITCH REVERSAL: 1, 1.5, or 2.5 vs. 0 or −1) and subsequently calculated the proportion of responses from each child that adhered to predictions. Dichotomizing the two principles describing proximity (PROXIMITY from the I-R model, PITCH PROXIMITY from the two-factor model) in this manner meant that they were identical.

For each variable, we also calculated the number of responses that would fulfill predictions if the first tone of responses was chosen randomly from the set of 25 possible responses. Each predictor variable was significant for the majority of individual children in each group, which is consistent with the significant simple associations observed with the group data (see Table 7). Moreover, one-sample *t*-tests confirmed that the mean percentage for each group was significantly higher than chance levels for each variable. The greatest number of children exhibited significant effects for INTERVALLIC DIFFERENCE (42 of 45), PROXIMITY/PITCH PROXIMITY (38 of 45), and PITCH REVERSAL (40 of 45), which is also consistent with the group data.[7] By contrast, REGISTRAL DIRECTION (15 of 45), REGISTRAL RETURN (24 of 45), and CLOSURE (21 of 45) were less consistent across individuals. In sum, response patterns from individual children mirrored those observed with the group data.

General Discussion

We examined melodic expectancies among listeners who varied in age. In Experiment 1, adults, 11-year-olds, and 8-year-olds rated how well various test tones continued fragments of melodies. In Experiment 2, 11-, 8-, and 5-year-olds sang continuations to two-tone stimulus intervals. One of our objectives was to compare two models of melodic expectancy, namely the I-R model (Narmour, 1990) and its simplified counterpart, the two-factor model (Schellenberg, 1997). We also sought to examine how expectancies change over development. A related goal was to determine whether melodic expectancies can be described using general perceptual principles that extend across domains and modalities.

Model Comparison

Following Cutting et al. (1992), we compared the I-R and two-factor models on the basis of their simplicity, scope, and selectivity. First, simplicity is defined by the number of parameters (e.g., regressors or predictors) required to explain data. The two-factor model is simpler than the original I-R model because it has three fewer predictor variables. Second, although good psychological models should generalize widely, the scope of some models with multiple parameters is too broad, allowing for successful prediction of random data (see Cutting et al., 1992). Neither the I-R nor the two-factor model predicted random data successfully. Finally, models should be selective, explaining patterned data (i.e., data they are supposed to explain) better than random data. For all age groups tested in both experiments, the two-factor model equaled or surpassed the I-R model in explanatory accuracy. In sum, the two models have equal scope, but the two-factor model is simpler and more selective.

Initial indications that the I-R model is needlessly complex were provided by Schellenberg's (1996, 1997) reanalyses of preexisting data. The model's overspecification was evident across a range of adult listeners, stimulus materials, and methods. The present results extend these earlier findings to predictions made in advance, to new sets of data, and to participants who vary in age, including children as young as 5 years of age. In short, the I-R model can be simplified without loss of explanatory accuracy, and evidence for this claim generalizes widely.

Did our experiments constitute fair tests of the I-R model? Might the model's redundancy be useful in other melodic contexts? Although neither of these questions can be answered definitively, recall that children's responses in Experiment 2 were not constrained in any way. Moreover, the melodies used as stimuli in Experiment 1 were "real" melodies that were not particularly simple. Although they were clearly tonal—as are the vast majority of melodies in Western classical, folk, and popular repertoires—they did not have the repeated, short motives that characterize many nursery tunes and popular songs (e.g., "Three Blind Mice," "Billie Jean"). Indeed, one could argue that they were more complex than some classic melodies (e.g., "Ode to Joy," from Beethoven's Ninth Symphony).

Other researchers have suggested that attempts to simplify the I-R model are rash (Krumhansl et al., 1999, 2000) or that simplified versions of the model are inferior in predictive accuracy (Thompson et al., 1997). These claims are at odds not only with the

---

[7] INTERVALLIC DIFFERENCE appears to be relatively important in these analyses and in the univariate analyses reported in Table 7. These findings are due to its overlap with PROXIMITY (Table 3). When both variables were tested simultaneously in the multivariate analyses (Table 8), the unique contribution of INTERVALLIC DIFFERENCE was weak, as it was in Experiment 1 (Table 5).

Table 9
Summary of Analyses of Individual Children From Experiment 2

Group	Registral direction	Intervallic difference	Registral return	Closure	Pitch proximity	Pitch reversal
Chance	47.81	44.00	17.52	64.00	52.00	33.14
11-year-olds						
M	61.43	76.83	35.87	77.94	82.54	62.06
SD	7.70	9.67	14.63	9.67	12.06	9.84
Min.	47.62	59.52	9.52	57.14	57.14	38.10
Max.	73.81	92.86	57.14	95.24	100.00	73.81
8-year-olds						
M	61.38	73.40	40.83	80.08	78.18	62.54
SD	8.89	11.21	23.17	10.31	13.57	15.38
Min.	47.62	52.38	11.90	66.67	52.38	38.10
Max.	76.19	95.12	92.86	97.56	97.62	92.86
5-year-olds						
M	58.40	79.30	31.41	76.52	85.51	56.98
SD	6.71	14.48	21.44	12.03	14.89	18.12
Min.	46.34	67.14	2.44	48.78	58.54	19.51
Max.	69.73	97.62	75.61	92.86	100.00	83.33

Note. Each predictor variable from the implication-realization (I-R) and two-factor models was dichotomized (proximity from the I-R model and pitch proximity from the two-factor model were identical). For each child, we calculated the percentage of responses that was consistent with predictions, separately for each variable. The table reports the mean, lowest (Min.), and highest (Max.) percentage, plus the standard deviation. In each case, the mean percentage was significantly higher than chance levels of responding (11-year-olds: $ps < .001$; 8-year-olds; $ps < .002$, 5-year-olds: $ps < .03$).

present findings but also with other direct comparisons of the I-R and two-factor models that have been conducted to date (Schellenberg, 1997). Such discrepancies can be explained, however, with relative ease. In one experiment (Thompson et al., 1997), musically trained participants used a keyboard to perform continuations to two-tone stimuli. Simplified versions of the principles from the I-R model were not as successful as the original principles at explaining responses, but the authors did not provide a direct comparison of the I-R and two-factor models. When we reanalyzed their data (Thompson et al., 1997, p. 1076), we found that the models were equivalent in their ability to predict the first tone of participants' productions.[8] In other experiments, Krumhansl et al. (1999, 2000) required participants to rate how well test tones continued melodic fragments, as in Experiment 1. Unfortunately, none of the fragments ended in an unambiguously large and implicative interval, which made the stimuli inadequate for testing either the I-R or the two-factor model.[9]

Our findings also provide evidence that the two-factor model is more sensitive than the I-R model to developmental changes in melodic expectancies. For the two-factor model, interpretable interactions between age and the model's predictor variables were evident in both experiments. The model's two predictors are grounded in the gestalt principle of proximity, with PITCH PROXIMITY instantiated at the simplest possible level (more proximate = more expected). PITCH REVERSAL is more complex, requiring consideration of the previous two tones a listener has heard; it describes expectancies for reversals in pitch direction after proximity has been violated (after a large implicative interval) and for tones that result in mirror-symmetric patterns (after small or large implicative intervals), which occur when the expected tone is proximate to the penultimate tone the listener has heard. The two-factor model appears to capture the distinction between simple and complex processes used in the formation of melodic expectancies.

By contrast, the redundancy and overspecificity of the five principles from the I-R model appear to blur distinctions between simple and complex processes, making the model relatively insensitive to developmental changes.

The two-factor model's distinction between first-order (simple) and second-order (complex) principles motivated two predictions: (a) Expectancies would be more strongly associated with PITCH PROXIMITY than with PITCH REVERSAL across development, and (b) PITCH REVERSAL would take longer than PITCH PROXIMITY to become fully operative. Overall, the results supported our hypotheses. PITCH PROXIMITY was associated with the outcome measure for all age groups in both experiments, and in all cases, such associations

---

[8] The data from Thompson et al. (1997, p. 1076, Appendix) were reanalyzed to compare the I-R and two-factor models directly. Two covariates were included with each model, as in the original analysis (see Thompson et al., 1997, p. 1073). After adjusting for bias, the difference between models in the deviance they explained was 2% or less for each comparison (i.e., participants with high musical training, low training, or the groups combined). In each case, the I-R model outperformed the two-factor model, but the difference fell well short of statistical significance (all $Fs < 1.1$).

[9] Seven of eight melodic fragments used by Krumhansl et al. (1999) ended in small intervals (0–4 semitones). The eighth ended with a perfect fifth (7 semitones). Strictly speaking, Narmour (1990) defined a perfect fifth as large, but he also considered it to be relatively ambiguous in terms of size because it is adjacent to the threshold between small and large intervals (6 semitones). Regardless, the fifth in question was not *open* (implicative), because the second tone fell on a stronger metrical position (on the downbeat of a measure) compared with the first tone (on the upbeat). In a subsequent experiment, Krumhansl et al. (2000) used a different set of eight melodic fragments, none of which ended in a large interval.

(simple and partial) were stronger than those observed for PITCH REVERSAL (and for any predictor from the I-R model). Others have documented the predominance of proximate tones in melodies from a variety of cultures (Dowling & Harwood, 1986; Huron, 2001). Our results are notable for clarifying that proximity is an important sequential grouping principle across a wide range of age levels.

Although PITCH PROXIMITY was equally strong across age groups in Experiment 2, it became stronger with age in Experiment 1. This discrepancy between experiments raises the possibility that the interaction with age in Experiment 1 could be task specific, possibly due to age-related differences in attending consistently to the task of making multiple ratings.[10] Although we attempted to compensate for this possibility by reducing the number of trials for younger children, any task that requires sustained attention is likely to reveal stronger associations among older participants, who have increased attentional capacity (Ruff & Lawson, 1990). By contrast, the interactive and creative aspects of the procedure in Experiment 2 appeared to engage most children throughout the test session. Regardless, in both experiments, PITCH PROXIMITY was a better predictor of response patterns for the youngest age group than PITCH REVERSAL was for the oldest group.

The results from Experiment 2 are particularly interesting in light of our inclusion of the ratio index variable in the expectancy models. This index for quantifying frequency ratios (Schellenberg & Trehub, 1994a) is culture free, as are the two principles in the two-factor model. Nonetheless, an additive combination of these three variables explained over half of the variance for each age group. Although it remains to be determined whether this finding would extend to children from non-Western musical cultures, it raises an interesting question that future research could address. It also suggests that claims of cultural relativism for musical structure and for music listening and performing (e.g., Blacking, 1992; Ellingson, 1992; Howe et al., 1998; Serafine, 1983) may be overstated.

Indeed, a wide variety of empirical evidence indicates that much of musical experience shows notable influences of perceptual and cognitive predispositions (for reviews, see Thompson & Schellenberg, 2002; Trehub, 2000; Trehub, Schellenberg, & Hill, 1997). The present findings are consistent with this perspective. In the introduction, we asked how music listening could possibly be "innately guided" when musical systems and structures vary widely both across and within cultures. In response, we speculate that perceptual and cognitive predispositions constrain the forms music can take and still be intelligible to the average listener. Within those constraints, however, much variety is evident. Perceptual grouping on the basis of proximity is likely to be one such predisposition that influences the structure of melodies, and of music in general (Huron, 2001). Other natural influences on musical pitch structures include processing biases that favor scales with unequal-sized steps (Trehub et al., 1999), intervals with small-integer frequency ratios (Schellenberg & Trehub, 1996a, 1996b; Trainor, 1997), and sequences with repeated tones (Schellenberg & Trehub, 1999). Melodies that exhibit these properties are easier to remember than other melodies, even for infants in the first year of life (Schellenberg & Trehub, 1996b, 1999; Trehub et al., 1999).

Developmental Changes

In line with our predictions, PITCH REVERSAL was underdeveloped among 8- and 11-year-olds compared with adults in Experiment 1, and among 5-year-olds compared with older children in Experiment 2. How can we account for this developmental shift in melodic expectancies? One possibility is that it is the consequence of maturity and increased exposure to music. Another possibility is that it is a by-product of general developmental progressions in perception and cognition that exert influence across domains and modalities. A third possibility is that learning and exposure in one auditory domain (i.e., speech) directly influences expectancies in another auditory domain (i.e., music).

For example, young children's failure to expect pitch reversals after large intervals in melodies could be a consequence of the frequent and large pitch shifts that they hear in speech, many of which are not followed by changes in direction. Adults and older children from a variety of cultures modify their way of speaking when addressing infants or younger children (e.g., Fernald & Simon, 1984; Fernald et al., 1989; Grieser & Kuhl, 1988). Compared with adult-directed speech, infant-directed speech is produced with higher pitch, exaggerated pitch contours (i.e., upward and downward shifts in pitch), slower tempo, shorter utterances, and extensive repetition (e.g., Ferguson, 1964; Fernald & Mazzie, 1991). Speech directed to toddlers and young children has similar modifications, although these become attenuated as the children become older (Garnica, 1977; Snow, 1972; Stern, Spieker, Barnett, & MacKain, 1983; Warren-Leubecker & Bohannon, 1984). Such modifications are identical to those made when adults speak to one another in an emotional manner (Trainor, Austin, & Desjardins, 2000). More important, the pitch excursions found in much of speech directed toward young children and in emotional adult-directed speech are frequently followed by further pitch movement in the same direction (Trainor et al., 2000). The short duration of many utterances to young children also ensures that they frequently end with relatively large pitch excursions. Thus, the relatively late emergence of PITCH REVERSAL could stem from early exposure to infant- and child-directed speech.

The relatively late importance of PITCH REVERSAL could also be related to general processes that are relevant to perceptual and cognitive development across domains, such as perceptual differentiation or improvements in working and sensory memory. Gibson's (1969) theory of perceptual development posits that increases in age and maturity are accompanied by greater differentiation. Perceptual learning is described as successive improvements "in the ability to extract information from the environment" (Gibson, 1969, p. 3). Gibson believed that ecologically valid stimuli are initially perceived as gestalts, or wholes. With exposure and learning, perceivers attend more and more to stimulus details.

---

[10] Another possibility is that the difference between experiments was due to the different set of stimulus intervals. When we limited the analysis of Experiment 2 to four implicative intervals (i.e., those used in Experiment 1: 2, 3, 9, and 10 semitones upward), however, response patterns were similar to those observed in the complete set of data. PITCH PROXIMITY did not interact with age. PITCH REVERSAL accounted for 14.2% and 18.7% of the variance among 11- and 8-year-olds, respectively, but for only 3.4% of the variance among 5-year-olds.

Age-related increases in differentiation are observed in speech perception (for a review, see Jusczyk, Houston, & Goodman, 1998), as they are in music perception. Young infants use prosodic cues in infant-directed speech to segment the speech stream into clauses (Hirsh-Pasek et al., 1987), much like the way they use pitch contour and tone duration to segment music into musical phrases (Jusczyk & Krumhansl, 1993). With increasing age, prosodic cues are used to segment speech into smaller units, such as noun and verb phrases (Jusczyk et al., 1992). At the word level, younger infants attend to entire syllables in speech (Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995). Older infants demonstrate increased sensitivity to smaller syllabic segments, such as allophonic cues to word boundaries (Jusczyk, Hohne, & Bauman, 1999).

In the musical domain, young infants are sensitive primarily to the global shape of melodies (Trehub et al., 1997), which is consistent with Gibson's (1969) theory. Infants perceive and remember melodies in a coarse-grained holistic manner, focusing on pitch contour (e.g., up, up, down, down) rather than the exact intervals between tones or the tones' exact pitch levels. (To illustrate, imagine the first three tones of "Three Blind Mice" and "The Star-Spangled Banner": They have an identical contour—down, down—but different intervals between tones.) When 8-month-olds are presented with a repeating *standard* melody, they readily notice when a *comparison* melody with a different contour is substituted for the standard (Trehub, Bull, & Thorpe, 1984). By contrast, they find it more difficult to detect a comparison melody if it is the same as the standard but presented in a different key (i.e., simply transposed upward or downward in pitch) or if it has the same contour but different intervals between tones. Similar processing styles are evident among children 4 to 6 years of age, who find contour-violating changes to a melody much easier to detect than contour-preserving changes (Morrongiello, Trehub, Thorpe, & Capodilupo, 1985; Pick et al., 1988). After hearing a melody for the first time, even adults are more likely to remember its contour than its exact intervals or pitch levels (Bartlett & Dowling, 1980; Dowling, 1978).

In general, however, melodic processing becomes more detailed with increases in age and exposure to music. For example, 4- to 6-year-olds notice when familiar melodies (e.g., "Happy Birthday") are performed out-of-tune (i.e., with some incorrect intervals between tones), although their performance improves if the out-of-tune melody has an altered contour in addition to some altered intervals (Trehub, Morrongiello, & Thorpe, 1985). With increasing age, intervallic but contour-preserving changes are noticed with greater accuracy (Andrews & Dowling, 1991). In fact, adults' ability to detect such changes in familiar melodies is not reliably different from contour-violating changes (Trehub et al., 1985). Adults are also better than 5-year-olds at detecting intervallic but contour-preserving changes to *unfamiliar* sequences of pure tones (Schellenberg & Trehub, 1999). In short, adults' representations of familiar and unfamiliar tunes are more detailed and exact than those of children. Our results suggest that melodic expectancies follow a similar developmental trajectory, becoming more detailed and specific as listeners increase in age.

Other aspects of cognitive development would contribute to the age-related changes in expectancies we observed. For example, working memory develops throughout childhood (Cowan, 1997). Increases in age are accompanied by increases in memory capacity

(Dempster, 1981) and speed (Kail, 1991), improved efficiency (Case, Kurland, & Goldberg, 1982), and greater specialization between verbal and spatial domains (Hale, Bronik, & Fry, 1997). In adulthood, the limited capacity of working memory is thought to constrain the number of different tones used in scales from around the world to a maximum of seven per octave (Dowling & Harwood, 1986). Consequently, many listeners find it difficult to understand and appreciate modern "art" music, which is often composed with scales that exceed this limit (e.g., 12-tone or microtonal scales). Improvements in working memory would increase the likelihood that listeners perceive, remember, and expect structures that extend over longer temporal spans, such as those described by PITCH REVERSAL (i.e., mirror-symmetric patterns, changes in direction). Developmental increases in the capacity of auditory sensory memory (Cowan, Nugent, Elliott, Ponomarev, & Saults, 1999; Saults & Cowan, 1996) would also make a contribution. For example, memory for a tone's pitch decays faster for 6- and 7-year-olds than for 10- to 12-year-olds, who, in turn, exhibit faster decay compared with adults (Keller & Cowan, 1994). Thus, developmental differences in retaining pitch in working or sensory memory could have contributed to the late emergence of expectancies for pitch reversals that we observed. This hypothesis could be tested in the future by examining whether measures of children's sensory and working memory abilities predict individual differences in the strength of the PITCH REVERSAL principle.

In sum, melody perception and cognition—and the expectancies that are formed—appear to be influenced by general developmental trends, including greater perceptual differentiation and improvements in memory. Melodic perception shifts from global processing based primarily on overall shape (contour), to local processing that includes specific pitch intervals between adjacent tones, to even more detailed local processing that considers three or more consecutive tones. Accordingly, expectancies based on pitch proximity appear relatively early in development, whereas expectancies based on mirror symmetry or on violations of proximity emerge later. The developmental differences we observed between the first-order (simple) and second-order (complex) principles of the two-factor model are similar to changes observed in other domains. For example, perspective-taking or "theory-of-mind" abilities improve during early childhood, moving from basic egocentrism (Piaget & Inhelder, 1956) to simple awareness of others' mental states ("What does X think?"; Wimmer & Perner, 1983) to more complex awareness ("What does X think Y is thinking?"; Perner & Wimmer, 1985). With increasing age and maturity, children also become more likely to succeed on tasks that require them to use embedded rules (e.g., "If this is the color game then the blue ones go on the right") rather than simple rules (e.g., "The blue ones go on the right"; Frye, Zelazo, & Burack, 1998; Zelazo & Frye, 1998).

Exposure to music and learning would also play an important role in the formation of melodic expectancies. With greater exposure, listeners would learn (implicitly in most cases) more detailed aspects of specific melodies and of melodies in general. Specifically, they would become familiar rather rapidly with the fact that small intervals (proximate tones) are prevalent in virtually all of the music they hear (Huron, 2001), such that expectancies for proximate tones would be evident early in development. With additional learning and exposure to music, violations of proximity (i.e., large intervals in a melody) would become more salient, and

more complex structures (mirror-symmetric patterns) would be perceived and remembered. On the one hand, the prevalence of small intervals in melodies is likely to be due to a domain-general predisposition to group stimuli on the basis of proximity. On the other hand, this predisposition would guarantee that proximate intervals predominate in the melodies people hear. Hence, effects of nature (a predisposition to group on the basis of proximity) and nurture (exposure to proximate stimuli) would be perfectly confounded, making it difficult if not impossible to tease them apart.

Expectancies, Memory, and Emotion

If memory capacity affects melodic expectancies, would melodic expectancies influence memory for melodies? Indeed, melodies that are judged to be consistent with adults' expectancies are better remembered than unexpected melodies (Schmuckler, 1997). Moreover, when listeners are asked to judge similarities between melodies, unexpected melodies are considered more similar to expected melodies than the reverse comparisons (Schmuckler, 1997). Schmuckler suggested that listeners have better memories for schematically central events (i.e., expected melodies) compared with peripheral events (i.e., unexpected melodies). When violations of expectancies are particularly salient, however, unexpected or incongruent music may actually be remembered with heightened accuracy (Boltz, Schulkind, & Kantra, 1991).

Melodic expectancies would also influence emotional reactions to music. Meyer (1956) proposed that listeners experience a degree of "surprise" when musical expectancies are violated rather than fulfilled (see Besson & Faïta, 1995, for neurological evidence), which, in turn, triggers emotional reactions. Many of the unexpected events in music that give rise to affective responding result from violations of learned, style-specific stimulus expectancies. But listeners are also likely to be surprised, at least at a subconscious level, by violations of either the PITCH PROXIMITY principle (e.g., when a large leap occurs in a melody) or the PITCH REVERSAL principle (e.g., when a large leap is not followed by a reversal in direction). The existence of such low-level processes helps to explain why listeners form expectancies and experience surprise when they hear an unfamiliar sounding melody from a foreign culture, and why they often continue to experience surprise or arousal at the same point in a musical piece after repeated listening (Gaver & Mandler, 1987). This perspective leads to another prediction that could be tested in the future: Compared with violations of PITCH REVERSAL, violations of PITCH PROXIMITY should be more strongly associated with surprise, arousal, and emotional responding, and such associations should be evident earlier in development.

It is important to clarify that our findings provide no evidence that listeners *prefer* or *like* what they expect. In fact, denials of expectancy may be central to the enjoyment of music (Gaver & Mandler, 1987), such that compositions are most enjoyed when they generate expectancies that are fulfilled and denied in a creative and aesthetically pleasing manner. According to Gaver and Mandler (1987), tunes that conform completely to expectancies are unlikely to cause arousal and therefore have little or no potential for emotional responding, a position consistent with Meyer (1956). By contrast, discrepancies between what listeners expect and what they actually hear are arousing, leading to emotional responding. "Mere exposure" is also relevant, because pieces tend to be liked simply because they have been heard before (Krugman, 1943; Mull, 1957). Nonetheless, many listeners report growing tired of music they have heard repeatedly. Laboratory findings confirm that listeners provide lower enjoyment ratings as tunes are presented repeatedly, particularly for tunes with simple structure (Heyduk, 1975). When music contains no elements of surprise because of overfamiliarity, listeners may experience satiation, similar to the way people satiate to particular foods (Raynor & Epstein, 2001), words (Balota & Black, 1997; Kounios, Kotz, & Holcomb, 2000), or faces (Lewis & Ellis, 2000). When hearing a piece for the first time, it may be interesting and enjoyable precisely because it violates expectancies generated either from perceptual predispositions (e.g., proximate tones) or from learned, style-specific factors (e.g., Western harmonic progressions, typical pop-song structures).

Limitations

The limitations of the present study are similar to those of others that have examined the validity of the I-R model. In each case, pitch was manipulated as an independent variable while differences in duration (rhythm) were held constant. Jones and her colleagues (Barnes & Jones, 2000; Jones, 1976, 1988; Jones & Boltz, 1989; Large & Jones, 1999) have emphasized that expectancies are *temporal*. Listeners have expectancies not only for specific musical events (notes, chords, etc.) but also for when these events will occur. Moreover, pitch and timing (i.e., rhythm and meter) often interact in their influence on the perception and cognition of melodies. For example, time manipulations affect the perception of pitch, and pitch manipulations affect the perception of time (Boltz, 1989a, 1989b, 1991, 1992, 1993; Crowder & Neath, 1995; Drake, Dowling, & Palmer, 1991; Jones, Boltz, & Kidd, 1982; Jones, Moynihan, MacKenzie, & Puente, 2002; Jones & Ralston, 1991; Jones, Summerell, & Marshburn, 1987; Kidd, Boltz, & Jones, 1984; Schmuckler & Boltz, 1994). Pitch and rhythm also interact in their influence on listeners' interpretation of the emotional valence of melodies (Schellenberg, Krysciak, & Campbell, 2000), and there is some evidence that musical expectancies are influenced more by rhythm structures than by pitch structures (Boltz, 1993; Schmuckler, 1990).

Thus, joint incorporation of rhythm and pitch as independent variables in tests of melodic expectancies is bound to yield new insights, insights that would result in further modifications and improvements to the quantified models tested in the present study. In Experiment 2, our stimulus intervals consisted of tones of equal duration and metrical emphasis. In Experiment 1, the final two tones of the stimulus melodies always conformed to a long–short (duration), strong–weak (meter), and stable–unstable (tonality) pattern. Future research could systematically vary each of these factors to determine their influence on expectancies. According to Narmour (1990), expectancies would become less specific and systematic if these patterns were altered or reversed.

Closer consideration of higher level learned factors would undoubtedly reveal additional complexities underlying the formation of musical expectancies. For example, melodies imply harmonies, and such implications are known to develop during childhood (Trainor & Trehub, 1992, 1994). One would therefore expect the influence of implied harmonies on melodic expectancies to become stronger with increases in age and exposure to music. Mu-

sical expectancies are further complicated by exposure to specific pieces (i.e., veridical expectancies). In some instances, familiar melodies appear to be processed more globally and schematically than unfamiliar melodies; in other instances, listening to familiar melodies is more analytical than it is with unfamiliar melodies (DeWitt & Samuel, 1990). Structural properties of familiar songs also affect melodic expectancies in childhood (Adachi, 1995), and memory for melodies presented recently affects subsequent melodic expectancies, even when listeners show no evidence of explicit memory for the melodies (Thompson, Balkwill, & Vernescu, 2000). In short, it is clear that melodic expectancies are extremely complex psychological phenomena. Nonetheless, the growth of scientific knowledge depends on limited and easily falsifiable theories (Popper, 1968). We believe that the two-factor model represents a good example of one such theory.

Final Remarks

In closing, we want to stress that Narmour's (1990) I-R model represents an important theoretical contribution to the field of music perception and cognition. His theory has generated much empirical research, including the behavioral studies central to the present report, as well as studies that examined associations between the I-R model and musical structure (Thompson & Stainton, 1995/1996, 1998). It is also important to reiterate that the two-factor model represents a simplification of the original I-R model. In fact, the two-factor model is based on specific aspects of the original model (expectancies for proximate tones, for pitch reversals after large intervals, and for mirror-symmetric patterns). Nonetheless, our results provide no support for Narmour's (1990) claim that the principles posited by the I-R model are innate and hardwired. None of the principles in its original form is retained in the simplified two-factor model, and our findings indicate that melodic expectancies change over development. Our results also provide unequivocal evidence that the I-R model can be simplified without loss of explanatory accuracy.

There are parallels between Narmour's contribution to the psychology of music and Chomsky's contribution to psycholinguistics. Both men are primarily theorists rather than empiricists, and both claim that innate factors constrain and predetermine processes of human communication. It remains to be determined whether Narmour's contribution to the psychology of music will prove to be as influential as Chomsky's contribution to psycholinguistics. Whereas Chomsky proposes a "universal grammar" for language, Narmour, through his I-R model, specifies a universal grammar for melodies (i.e., innately specified rules governing the order of tones in melodies). The present findings are generally consistent with the spirit of Narmour's proposal, yet they also imply that the rules of the grammar are considerably simpler than those specified by the I-R model. Indeed, our results imply that melodic expectancies are neither esoteric nor wedded to music. Rather, the basic rules governing melodic expectancies and their development appear to stem from perceptual and cognitive biases that extend beyond music in particular and audition in general.

References

Adachi, M. (1995). *Tracking the nature of melodic expectancy development in musical children.* Unpublished doctoral dissertation, University of Washington.

Adachi, M., & Carlsen, J. C. (1995/1996). Measuring melodic expectancies with children. *Bulletin of the Council for Research in Music Education, 127,* 1–7.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Andrews, M. W., & Dowling, W. J. (1991). The development of perception of interleaved melodies and control of auditory attention. *Music Perception, 8,* 349–368.

Baillargeon, R. (1999). Young infants' expectations about hidden objects: A reply to three challenges. *Developmental Science, 2,* 115–132.

Balota, D. A., & Black, S. (1997). Semantic satiation in healthy young and older adults. *Memory & Cognition, 25,* 190–202.

Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology, 41,* 254–311.

Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 501–515.

Besson, M., & Faita, F. (1995). An event-related potential (ERP) study of musical expectancy: Comparisons of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1278–1296.

Bharucha, J. J. (1994). Tonality and expectation. In R. Aiello & J. Sloboda (Eds.), *Musical perceptions* (pp. 213–239). New York: Oxford University Press.

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

Bigand, E., Madurell, F., Tillmann, B., & Pineau, M. (1999). Effect of global structure and temporal organization on chord processing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 184–197.

Bigand, E., & Pineau, M. (1997). Global context effects on musical expectancy. *Perception & Psychophysics, 59,* 1098–1107.

Blacking, J. (1992). The biology of music-making. In H. Myers (Ed.), *Ethnomusicology: An introduction* (pp. 301–314). New York: Norton.

Boltz, M. (1989a). Perceiving the end: Effects of tonal relationships on melodic completion. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 749–761.

Boltz, M. (1989b). Rhythm and "good endings": Effects of temporal structure on tonality judgments. *Perception & Psychophysics, 46,* 9–17.

Boltz, M. (1991). Some structural determinants of melody recall. *Memory & Cognition, 19,* 239–251.

Boltz, M. (1992). The remembering of auditory event durations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 938–956.

Boltz, M. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception & Psychophysics, 53,* 585–600.

Boltz, M., Schulkind, M., & Kantra, S. (1991). Effects of background music on the remembering of filmed events. *Memory & Cognition, 19,* 593–606.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Carlsen, J. C. (1981). Some factors which influence melodic expectancy. *Psychomusicology, 1,* 12–29.

Case, R., Kurland, M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33,* 386–404.

Chaplin, J. P. (1985). *Dictionary of psychology* (2nd ed., rev.). New York: Laurel.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments, & Computers, 25,* 257–271.

Cowan, N. (1997). The development of working memory. In N. Cowan

(Ed.), *The development of memory in childhood* (pp. 163–199). East Sussex, United Kingdom: Psychology Press.

Cowan, N., Nugent, L. D., Elliott, E. M., Ponomarev, I., & Saults, J. S. (1999). The role of attention in the development of short-term memory: Age differences in the verbal span of apprehension. *Child Development, 70,* 1082–1097.

Crowder, R. G., & Neath, I. (1995). The influence of pitch on time perception in short melodies. *Music Perception, 12,* 379–386.

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

Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception & Psychophysics, 57,* 451–462.

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.

Davidson, L. (1985). Tonal structures of children's early songs. *Music Perception, 2,* 361–374.

Dempster, F. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin, 89,* 63–100.

Dennett, D. C. (1991). *Consciousness explained.* Boston: Little, Brown.

DeWitt, L. A., & Samuel, A. G. (1990). The role of knowledge-based expectations in music perception: Evidence from musical restoration. *Journal of Experimental Psychology: General, 119,* 123–144.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology, 5,* 322–337.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review, 85,* 341–354.

Dowling, W. J. (1984). Development of musical schemata in children's spontaneous singing. In W. R. Crozier & A. J. Chapman (Eds.), *Cognitive processes in the perception of art* (pp. 145–163). Amsterdam, the Netherlands: Elsevier.

Dowling, W. J. (1990). Expectancy and attention in melody perception. *Psychomusicology, 9,* 148–160.

Dowling, W. J. (2001). Perception of music. In E. B. Goldstein (Ed.), *Blackwell handbook of perception* (pp. 469–498). Malden, MA: Blackwell.

Dowling, W. J., & Harwood, D. L. (1986). *Music cognition.* Orlando, FL: Academic Press.

Dowling, W. J., Lung, K. M.-T., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics, 41,* 642–656.

Drake, C., Dowling, W. J., & Palmer, C. (1991). Accent structures in the reproduction of simple tunes by children and adult pianists. *Music Perception, 8,* 315–334.

Drexler, E. N. (1938). A study of the development of the ability to carry a melody at the preschool level. *Child Development, 9,* 319–332.

Ellingson, T. (1992). Transcription. In H. Myers (Ed.), *Ethnomusicology: An introduction* (pp. 110–152). New York: Norton.

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist, 66,* 103–114.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology, 27,* 209–221.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20,* 104–113.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16,* 477–501.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

Frye, D., Zelazo, P. D., & Burack, J. A. (1998). Cognitive complexity and control: I. Theory of mind in typical and atypical development. *Current Directions in Psychological Science, 7,* 116–121.

Garnica, O. (1977). Some prosodic and paralinguistic features of speech to young children. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 63–88). Cambridge, United Kingdom: Cambridge University Press.

Gaver, W. W., & Mandler, G. (1987). Play it again, Sam: On liking music. *Cognition and Emotion, 1,* 259–282.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Appleton-Century-Croft.

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology, 24,* 14–20.

Hale, S., Bronik, M. D., & Fry, A. F. (1997). Verbal and spatial working memory in school-age children: Developmental differences in susceptibility to interference. *Developmental Psychology, 33,* 364–371.

Hargreaves, D. J. (1986). *The developmental psychology of music.* Cambridge, United Kingdom: Cambridge University Press.

Heyduk, R. G. (1975). Rated preference for musical compositions as it relates to complexity and exposure frequency. *Perception & Psychophysics, 17,* 84–90.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Wright Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26,* 269–286.

Howe, M. J. A., Davidson, J. W., & Sloboda, J. A. (1998). Innate talents: Reality or myth? *Behavioral and Brain Sciences, 21,* 399–442.

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception, 19,* 1–64.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83,* 323–355.

Jones, M. R. (1988). What does expectancy mean? *Behavioral and Brain Sciences, 11,* 387–388.

Jones, M. R. (1990). Learning and the development of expectancies: An interactionist approach. *Psychomusicology, 9,* 193–228.

Jones, M. R., & Boltz, M. G. (1989). Dynamic attending and responses to time. *Psychological Review, 96,* 459–491.

Jones, M. R., Boltz, M. G., & Kidd, G. (1982). Controlled attending as a function of melodic and temporal context. *Perception & Psychophysics, 32,* 211–218.

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science, 13,* 313–319.

Jones, M. R., & Ralston, J. T. (1991). Some influences of accent structure on melody recognition. *Memory & Cognition, 19,* 8–20.

Jones, M. R., Summerell, L., & Marshburn, E. (1987). Recognizing melodies: A dynamic interpretation. *Quarterly Journal of Experimental Psychology, 39A,* 89–121.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology, 24,* 252–293.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics, 61,* 1465–1476.

Jusczyk, P. W., Houston, D., & Goodman, M. (1998). Speech perception during the first year. In A. Slater (Ed.), *Perceptual development: Visual, auditory, and speech perception in infancy* (pp. 357–388). East Sussex, United Kingdom: Psychology Press.

Jusczyk, P. W., Juszczyk, A. M., Kennedy, L. J., Schomberg, T., & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 822–836.

Jusczyk, P. W., & Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of*

Experimental Psychology: Human Perception and Performance, 19, 627–640.

Kail, R. V. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin, 109,* 490–501.

Keller, T. A., & Cowan, N. (1994). Developmental increase in the duration of memory for tone pitch. *Developmental Psychology, 30,* 855–863.

Kidd, G., Boltz, M., & Jones, M. R. (1984). Some effects of rhythmic context on melody recognition. *American Journal of Psychology, 97,* 153–173.

Kirsh, I. (1999). Response expectancy: An introduction. In I. Kirsh (Ed.), *How expectancies shape experience* (pp. 3–13). Washington, DC: American Psychological Association.

Kounios, J., Kotz, S. A., & Holcomb, P. J. (2000). On the locus of the semantic satiation effect: Evidence from event-related brain potentials. *Memory & Cognition, 28,* 1366–1377.

Krugman, H. E. (1943). Affective responses to music as a function of familiarity. *Journal of Abnormal and Social Psychology, 38,* 388–392.

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

Krumhansl, C. L. (1995). Effects of musical context on similarity and expectancy. *Systematische Musikwissenschaft* [Systematic Musicology], *3,* 211–250.

Krumhansl, C. L. (1997). Effects of perceptual organization and musical form on melodic expectancies. In M. Leman (Ed.), *Music, gestalt, and computing: Studies in cognitive and systematic musicology* (pp. 294–320). Berlin: Springer-Verlag.

Krumhansl, C. L., & Keil, F. C. (1982). Acquisition of the hierarchy of tonal functions in music. *Memory & Cognition, 10,* 243–251.

Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic expectation in Finnish spiritual folk hymns: Convergence of statistical, behavioral, and computational approaches. *Music Perception, 17,* 151–195.

Krumhansl, C. L., Toivanen, P., Eerola, T., Toiviainen, P., Järvinen, T., & Louhivuori, J. (2000). Cross-cultural music cognition: Cognitive methodology applied to North Sami yoiks. *Cognition, 76,* 13–58.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How we track time varying events. *Psychological Review, 106,* 119–159.

Lewis, M. B., & Ellis, H. D. (2000). Satiation in name and face recognition. *Memory & Cognition, 28,* 783–788.

Lynch, M. P., Eilers, R. E., Oller, D. K., & Urbano, R. C. (1990). Innateness, experience, and music perception. *Psychological Science, 1,* 272–276.

Maddux, J. E. (1999). Expectancies and the social–cognitive perspective: Basic principles, processes, and variables. In I. Kirsh (Ed.), *How expectancies shape experience* (pp. 17–39). Washington, DC: American Psychological Association.

Meyer, L. B. (1956). *Emotion and meaning in music.* Chicago: University of Chicago Press.

Meyer, L. B. (1973). *Explaining music: Essays and explorations.* Chicago: University of Chicago Press.

Miller, L. B. (1987). Children's musical behaviors in the natural environment. In J. C. Perry, I. W. Perry, & T. W. Draper (Eds.), *Music and child development* (pp. 206–224). New York: Springer-Verlag.

Morrongiello, B. A. (1992). Effects of training on children's perception of music: A review. *Psychology of Music, 20,* 29–41.

Morrongiello, B. A., Trehub, S. E., Thorpe, L. A., & Capodilupo, S. (1985). Children's perception of melodies: The role of contour, frequency, and rate of presentation. *Journal of Experimental Child Psychology, 40,* 279–292.

Mull, H. K. (1957). The effect of repetition upon the enjoyment of modern music. *Journal of Psychology, 43,* 155–162.

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.* Chicago: University of Chicago Press.

Narmour, E. (1992). *The analysis and cognition of melodic complexity.* Chicago: University of Chicago Press.

Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 211–238). New York: Guilford Press.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that . . . ": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39,* 437–471.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space.* London: Routledge & Kegan Paul.

Pick, A. D., Palmer, C. F., Hennessy, B. L., Unze, M. G., Jones, R. K., & Richardson, R. M. (1988). Children's perception of certain musical properties: Scale and contour. *Journal of Experimental Child Psychology, 45,* 28–51.

Popper, K. R. (1968). *Conjectures and refutations: The growth of scientific knowledge.* New York: Harper & Row.

Raynor, H. A., & Epstein, L. H. (2001). Dietary variety, energy regulation, and obesity. *Psychological Bulletin, 127,* 325–341.

Ruff, H. A., & Lawson, K. R. (1990). Development of sustained, focused attention in young children during free play. *Developmental Psychology, 26,* 85–93.

Saults, J. S., & Cowan, N. (1996). The development of memory for ignored speech. *Journal of Experimental Child Psychology, 63,* 239–261.

Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realization model. *Cognition, 58,* 75–125.

Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception, 14,* 295–318.

Schellenberg, E. G. (2001, April). *Becoming an enculturated music listener.* Paper presented at the biennial meeting of the Society for Research in Child Development, Minneapolis, MN.

Schellenberg, E. G., Krysciak, A. M., & Campbell, R. J. (2000). Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception, 18,* 155–171.

Schellenberg, E. G., & Trainor, L. J. (1996). Sensory consonance and the perceptual similarity of complex-tone harmonic intervals: Tests of adult and infant listeners. *Journal of the Acoustical Society of America, 100,* 3321–3328.

Schellenberg, E. G., & Trehub, S. E. (1994a). Frequency ratios and the discrimination of pure tone sequences. *Perception & Psychophysics, 56,* 472–478.

Schellenberg, E. G., & Trehub, S. E. (1994b). Frequency ratios and the perception of tone patterns. *Psychonomic Bulletin & Review, 1,* 191–201.

Schellenberg, E. G., & Trehub, S. E. (1996a). Children's discrimination of melodic intervals. *Developmental Psychology, 32,* 1039–1050.

Schellenberg, E. G., & Trehub, S. E. (1996b). Natural musical intervals: Evidence from infant listeners. *Psychological Science, 7,* 272–277.

Schellenberg, E. G., & Trehub, S. E. (1999). Redundancy, conventionality, and the discrimination of tone sequences: A developmental perspective. *Journal of Experimental Child Psychology, 74,* 107–127.

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.

Schmuckler, M. A. (1997). Expectancy effects in memory for melodies. *Canadian Journal of Experimental Psychology, 51,* 292–305.

Schmuckler, M. A., & Boltz, M. G. (1994). Harmonic and rhythmic influences on musical expectancy. *Perception & Psychophysics, 56,* 313–325.

Serafine, M. L. (1983). Cognition in music. *Cognition, 14,* 119–183.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development, 43,* 549–565.

Spelke, E. S. (1994). Initial knowledge: Six suggestions. *Cognition, 50,* 431–445.

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language, 10,* 1–15.

Tekman, H. G., & Bharucha, J. J. (1992). Time course of chord priming. *Perception & Psychophysics, 51,* 33–39.

Thompson, W. F., Balkwill, L.-L., & Vernescu, R. (2000). Expectancies generated by recent exposure to melodic sequences. *Memory & Cognition, 28,* 547–555.

Thompson, W. F., Cuddy, L. L., & Plaus, C. (1997). Expectancies generated by melodic intervals: Evaluation of principles of melodic implication in a melody-completion task. *Perception & Psychophysics, 59,* 1069–1076.

Thompson, W. F., & Schellenberg, E. G. (2002). Cognitive constraints on music listening. In R. Colwell & C. Richardson (Eds.), *The new handbook of research on music teaching and learning* (pp. 461–486). New York: Oxford University Press.

Thompson, W. F., & Stainton, M. (1995/1996). Using humdrum to analyze melodic structure: An assessment of Narmour's implication-realization model. *Computing in Musicology, 10,* 24–33.

Thompson, W. F., & Stainton, M. (1998). Expectancy in Bohemian folk song melodies: Evaluation of implicative principles for implicative and closural intervals. *Music Perception, 15,* 231–252.

Tillmann, B., Bigand, E., & Pineau, M. (1998). Effects of global and local contexts on harmonic expectancy. *Music Perception, 16,* 99–117.

Trainor, L. J. (1997). The effect of frequency ratio on infants' and adults' discrimination of simultaneous intervals. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1427–1438.

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science, 11,* 188–195.

Trainor, L. J., & Heinmiller, B. M. (1998). The development of evaluative responses to music: Infants prefer to listen to consonance over dissonance. *Infant Behavior & Development, 21,* 77–88.

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.

Trainor, L. J., & Trehub, S. E. (1994). Key membership and implied harmony in Western tonal music: Developmental perspectives. *Perception & Psychophysics, 56,* 125–132.

Trehub, S. E. (2000). Human processing predispositions and musical universals. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 427–448). Cambridge, MA: MIT Press.

Trehub, S. E., Bull, D., & Thorpe, L. A. (1984). Infants' perception of melodies: The role of melodic contour. *Child Development, 55,* 821–830.

Trehub, S. E., Morrongiello, B. A., & Thorpe, L. A. (1985). Children's perception of familiar melodies: The role of intervals, contour, and key. *Psychomusicology, 5,* 39–48.

Trehub, S. E., Schellenberg, E. G., & Hill, D. (1997). The origins of music perception and cognition: A developmental perspective. In I. Deliège & J. A. Sloboda (Eds.), *Perception and cognition of music* (pp. 103–128). East Sussex, United Kingdom: Psychology Press.

Trehub, S. E., Schellenberg, E. G., & Kamenetsky, S. B. (1999). Infants' and adults' perception of scale structure. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 965–975.

Unyk, A. M., & Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology, 7,* 3–23.

von Hippel, P. (2000). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception, 17,* 315–327.

Warren-Leubecker, A., & Bohannon, J. N. (1984). Intonation patterns in child-directed speech: Mother–father differences. *Child Development, 55,* 1379–1385.

Wassum, S. (1979). Elementary school children's vocal range. *Journal of Research in Music Education, 27,* 214–226.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103–128.

Zelazo, P. D., & Frye, D. (1998). Cognitive complexity and control: II. The development of executive function in childhood. *Current Directions in Psychological Science, 7,* 121–126.

Zentner, M. R., & Kagan, J. (1996, September 5). Perception of music by infants. *Nature, 383,* 29.

Wanted: Your Old Issues!

As APA continues its efforts to digitize journal issues for the PsycARTICLES database, we are finding that older issues are increasingly unavailable in our inventory. We are turning to our long-time subscribers for assistance. If you would like to donate any back issues toward this effort (preceding 1982), please get in touch with us at journals@apa.org and specify the journal titles, volumes, and issue numbers that you would like us to take off your hands.