

DYNAMIC MELODIC EXPECTANCY

DISSERTATION

Presented in Partial Fulfillment of the Requirements  
for the Degree Doctor of Philosophy  
in the Graduate School of The Ohio State University

By

Bret J. Aarden, M.A.

The Ohio State University

2003

Dissertation Committee:

Professor David Huron, Adviser

Professor Patricia Flowers

Professor Mari Riess Jones

Professor Lee Potter

Approved by

---

Adviser

School of Music

## ABSTRACT

The most common method for measuring melodic expectancy is the “probe-tone” design, which relies on a retrospective report of expectancy. Here a direct measure of expectancy is introduced, one that uses a speeded, serial categorization task. An analysis of the reaction time data showed that “Implication-Realization” contour models of melodic expectancy provide a good fit. Further analysis suggests that some assumptions of these contour models may not be valid.

The traditional “key profile” model of tonality was not found to contribute significantly to the model. Following Krumhansl’s (1990) argument that tonality is learned from the statistical distribution of scale degrees, a tonality model based on the actual probability of scale degrees did significantly improve the fit of the model.

It is proposed that the probe-tone method for measuring key profiles encourages listeners to treat the probe tone as being in phrase-final position. Indeed, the key profile was found to be much more similar to the distribution of phrase-final notes than to the distribution of all melodic notes.

A second experiment measured reaction times to notes that subjects expected to be phrase-final. In this experiment the key profile contributed significantly to the fit of the model.

It is concluded that the probe-tone design creates a task demand to hear the tone as a phrase-final note, and the key profile reflects a learned sensitivity to the distribution of notes at ends of melodies. The “key profile” produced by the new reaction-time design is apparently related to the general distribution of notes in melodies. The results of this study indicate that the relationship between melodic structure and melodic expectation is more straightforward than has been previously demonstrated. Melodic expectation appears to be related directly to the structure and distribution of events in the music.

## ACKNOWLEDGMENTS

There are a number of people whose contributions to this dissertation need to be acknowledged. I would like to thank my advisor David Huron for his endless support and pragmatic advice over the course of this project. I believe his lab has been an unparalleled place to pursue the musicological study of psychology. Without Paul von Hippel's work at Ohio State the concept for these experiments would never have occurred to me. Many office and coffeehouse conversations with David Butler informed my thinking about the state of music psychology and are reflected in the grundgestalt of this work. Finally, I would especially like to thank my wife Althea for numerous conversations and pep talks, such as the one that led to chapter 5. I hope you can see your good influences reflected here.

## VITA

2002.....	Graduate Minor, Quantitative Psychology The Ohio State University
2001.....	M. A., Music Theory The Ohio State University
1998 – present.....	Dean’s Distinguished Fellow, Ohio State University
1998.....	B. A., New College of Florida

## PUBLICATIONS

### Research publications

1. Aarden, Bret. (2002). Expectancy vs. retrospective perception: Reconsidering the effects of schema and continuation judgments on measures of melodic expectancy. In *Proceedings of the 7<sup>th</sup> International Conference on Music Perception and Cognition*. Ed. C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick. Adelaide: Causal Productions.
2. Aarden, Bret and Huron, David. (2001). Mapping European folksong: Geographical localization of musical features. *Computing in Musicology*, **12**:169–183.

## FIELDS OF STUDY

Major Field: Music

## TABLE OF CONTENTS

Abstract.....	ii
Acknowledgments.....	iv
Vita.....	v
List of Tables .....	viii
List of Figures.....	ix
Chapter 1: Introduction.....	1
Chapter 2: Theories of Melodic Expectancy .....	6
The Grand Illusion: Harmonic Influences on Melody.....	7
An Overview and Critique of the Tonal Hierarchy Theory.....	11
The Static Key Profile.....	13
Ordering Through Time.....	15
Effects of Stimulus Structure .....	17
Equating Chord With Key .....	22
Problems Explaining the Key Profiles.....	25
An Overview and Critique of the Implication-Realization Model .....	27
The Gestalt principles .....	30
Redundancy.....	32
Regression to the Mean.....	35
Chapter 3: Methods of Measurement.....	37
Previous Methods for Measuring Melodic Expectancy.....	38
A New Method.....	39

Chapter 4: Expectancy for Continuation (Experiment 1) .....	42
Methods.....	43
Stimuli.....	43
Equipment.....	45
Subjects and Procedure.....	45
Results.....	48
Univariate Main Effects.....	48
Multivariate Analysis.....	53
Discussion.....	57
Overview.....	66
Chapter 5: Expectancy for Closure (Experiment 2).....	69
Methods.....	69
Stimuli.....	70
Subjects, Equipment, and Procedure .....	71
Results.....	72
Univariate Main Effects.....	72
Multivariate analysis.....	75
Discussion.....	79
Further evidence from key-finding.....	80
Conclusions.....	84
Chapter 6: Re-examining the Implication-Realization Model.....	86
The 5-Factor Parameters model .....	86
The Archetype Model .....	88
The Statistical Model .....	92
Overview.....	93
Chapter 7: Conclusions.....	95
References.....	100

## LIST OF TABLES

<u>Table 2.1.</u>	The Seven Prospective 3-Note Archetypes of Narmour's Implication-Realization Model .....	29
<u>Table 4.1.</u>	Regression Analysis of the I-R and Tonality Models on Log Reaction Time .....	56
<u>Table 4.2.</u>	Results of the I-R/Tonality Model Including the Probability Predictor .....	64
<u>Table 5.1.</u>	Results of the Regression Analysis of Experiment 2 .....	76
<u>Table 5.2.</u>	Performance of scale degree templates in key identification .....	83
<u>Table 6.1.</u>	Regression Analysis of the 5-Factor I-R Model of Krumhansl (1995) and Schellenberg (1996) .....	87
<u>Table 6.2.</u>	The Prospective Archetypes of Narmour's I-R Theory, Including Instance Counts .....	89
<u>Table 6.3.</u>	Regression Analysis of the 5-Factor I-R Archetype Model .....	91
<u>Table 6.4.</u>	Model Fit Criteria for the Four Variants of the I-R Model .....	94



## LIST OF FIGURES

<u>Figure 1.1.</u>	The major-key diatonic “tonal hierarchy,” as measured by Krumhansl and Kessler (1982). .....	3
<u>Figure 2.1.</u>	Two chord progressions in the key of A major. In (a) the tonic is stable when progressing from <i>V</i> to <i>I</i> , but in (b) the tonic is unstable when suspended from <i>IV</i> to <i>V</i> . .....	14
<u>Figure 2.2.</u>	The ascending major melodic scale pattern (from Butler, 1989a). .....	18
<u>Figure 2.3.</u>	The tone profile for Krumhansl & Shepard (1979) major melodic scale patterns, for subjects with moderate musical training (Group 2). .....	19
<u>Figure 2.4.</u>	The path of a diatonic harmonic progression ( <i>IV-V-vi-IV-I-vi-ii-V-I</i> ) through foreign key space (from Krumhansl & Kessler, 1982). .....	23
<u>Figure 4.1.</u>	Five of the 37 melodic phrases from the Essen Folksong Collection used as stimuli in Experiment 1. ....	44
<u>Figure 4.2.</u>	The observed reaction times at the various levels of the pitch-proximity variable. The expected trend of the variable is shown as a dotted line. ....	49

<u>Figure 4.3.</u>	The observed reaction times at the five levels of the pitch-reversal variable. The expected trend of the variable is shown as a dotted line. The category labels indicate the size of the first and second intervals; a ‘+’ indicates continuation, and ‘-’ indicates reversal.....	50
<u>Figure 4.4.</u>	The observed reaction times at the two levels of the process variable. The expected trend of the variable is shown as a dotted line.....	51
<u>Figure 4.5.</u>	The observed reaction times at the seven diatonic levels of the key-profile variable, ordered according to value (rather than scale degree). The expected trend of the variable is shown as a dotted line.....	52
<u>Figure 4.6.</u>	The observed reaction times at the two levels of the unison variable. The expected trend of the variable is shown as a dotted line.....	53
<u>Figure 4.7.</u>	Percentage accuracy plotted against reaction time. Reaction times were binned over intervals of 50 milliseconds.....	55
<u>Figure 4.8.</u>	Scale degree weight estimates (displayed on a reverse ordinate), and the key profile ratings ( $r = -0.53$ ). .....	59
<u>Figure 4.9.</u>	Zero-order probabilities of diatonic scale degree occurrence in dozens of Western art song melodies, as measured by Youngblood (1958), N = 2668, and Knopoff and Hutchinson (1983), N = 25,122.....	61
<u>Figure 4.10.</u>	The estimated zero-order probabilities of diatonic scale degrees for a major-key folksong from the Essen Folksong Collection. Values were averaged from 1000 folksongs, N = 49,265. The major key profile is shown for comparison, using a 7-point Likert scale on the right ordinate axis.....	63

<u>Figure 4.11.</u> Scale degree weight estimates (displayed on a reverse ordinate), plotted with the zero-order probability of scale degrees ( $r = -0.87$ ).	65
<u>Figure 4.12.</u> The estimated zero-order probabilities of phrase-final scale degrees for major-key folksongs from the Essen Folksong Collection. Values were averaged from the notes in 1000 folksongs, $N = 5832$ . The major key profile is included as a comparison ( $r = 0.87$ ).	67
<u>Figure 5.1.</u> Seven of the 82 melodic phrases used in Experiment 2. One melody is shown for each diatonic scale degree (SD) ending, 1 through 7.	71
<u>Figure 5.2.</u> The observed reaction times (solid line) at the various levels of the pitch-proximity variable. The expected trend of the variable is shown as a dotted line.	73
<u>Figure 5.3.</u> The observed reaction times at the two levels of the pitch-reversal variable. The expected trend of the variable is shown as a dotted line.	74
<u>Figure 5.4.</u> The observed reaction times at the two levels of the process variable. The expected trend of the variable is shown as a dotted line.	74
<u>Figure 5.5.</u> The observed reaction times at the various levels of the key-profile variable. The expected trend of the variable is shown as a dotted line.	75
<u>Figure 5.6.</u> Scale degree weight estimates in log reaction time (shown on a reverse ordinate axis), along with the major key profile ( $r = -0.91$ ).	78

Figure 5.7. The average distribution of scale degree durations for pieces from two samples. The monophonic sample consisted of 1000 major-key monophonic folksongs from the Essen Folksong Collection (49,265 notes), and the polyphonic sample contained 250 major-key polyphonic segments of movements from the CCARH MuseData database (81,524 notes). ..... 82

## CHAPTER 1

### INTRODUCTION

The study of the psychology of melodic expectancy has a history going back over a century. Studies by Theodor Lipps (1885/1926) and Max Meyer (1901) in the late nineteenth century supported the established wisdom that humans hear musical intervals in terms of simple integer frequency ratios. In the early twentieth century, a more comprehensive series of experiments by William Van Dyke Bingham made it clear that a more musical explanation was necessary (Bingham, 1910).

With the rise of behaviorism, however, cultural topics such as music fell out of favor in psychology. It took a musicologist, Leonard Meyer, to reopen the issue of the psychology of music in the 1950s. His book, *Emotion and Meaning in Music*, argued — in part — that the aesthetic appeal of listening to melodies involves forming expectations about what will happen, and having those expectations confirmed or denied (L. B. Meyer, 1956).

After the advent of the “cognitive revolution” in the 1960s, interest was reawakened in experimental music psychology. The melodic studies of the next decade generally

worked within the limits of existing music theoretic concepts such as contour, interval, scale, key, and transposition, without positing novel mental representations (Cuddy & Cohen, 1976; Deutsch, 1969; Dowling, 1971).

In 1979, Shepard and Krumhansl published a seminal paper measuring responses to individual notes. Their technique became known as the “probe tone” method, and it threw open the doors to the detailed study of melodic perception. In recent years, as the focus of cognitive science has broadened, Krumhansl has reinterpreted studies using this method as research into expectancy (Krumhansl, 1995).

The Shepard and Krumhansl probe-tone method has been very influential, and is possibly the single most famous technique for studying music perception in the psychological literature. By stopping the melody at specific places and asking listeners to rate the final note according to some criterion, researchers replicated earlier studies showing the importance of the distinction between in-key and out-of-key notes. In addition, it was shown that the members of the tonic chord are especially important, particularly the tonic itself.

According to Krumhansl (1990), this hierarchy of importance among scale degrees — dubbed the “tonal hierarchy” (see Figure 1.1) — is learned from long-term exposure to music. According to her theory, notes that occur most frequently are rated more highly in these experiments.

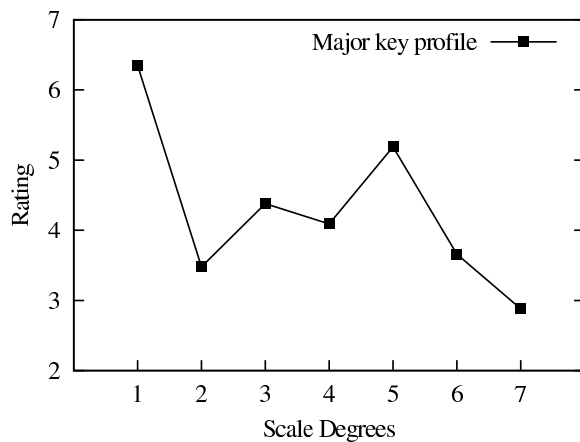


Figure 1.1. The major-key diatonic “tonal hierarchy,” as measured by Krumhansl and Kessler (1982).

In the early 1990s, music theorist Eugene Narmour introduced another influential theory of melodic expectancy, inspired by the work of his mentor, Leonard Meyer (Narmour, 1990, 1992). Dubbed the “Implication-Realization” (I-R) theory, it expanded on Meyer’s idea that some types of melodic expectancy are cross-cultural.

Narmour made the distinction between melodic contexts that create strong expectancy for particular continuations, and “closure” contexts that produce few or no ensuing expectations. A simple quantitative model of the I-R theory was developed and tested, and many parts of it were confirmed (Krumhansl, 1995; Schellenberg, 1996). Further work has both served to simplify the quantification of its principles (Schellenberg, 1996, 1997), and called into question its assumptions (von Hippel & Huron, 2000).

It is interesting to note that the I-R theory’s emphasis on closure may have important consequences for the probe-tone method. Endings are strongly linked with moments of

closure, and Narmour's theory states that expectations are different at closure. Because the probe-tone method requires stopping the melody before collecting responses, that may result in limiting the scope of expectancy being studied.

The topic of this document is the introduction of a new method for studying melodic expectancy (Chapter 4). The goal of this new design is to avoid inadvertently measuring closure and other cognitive factors that may not directly relate to expectation. Whereas the probe tone method takes measurements after the stimulus is finished, the method introduced here takes reaction-time measurements in the flow of music without stopping the melodies.

The results from this design are shown to be compatible with findings from previous studies, but there are some interesting differences. Two recent studies have identified a few I-R factors that explain melodic expectancy fairly well (Schellenberg, 1997; von Hippel, 2002). When used to predict reaction times in the new design, they each explain significant amounts of the variance in the data.

When the data are analyzed further, however, it appears that the current I-R models are over-specified (Chapter 6). Some assumptions made in developing the I-R models may impose restrictions that are not supported by the data.

More importantly, there are notable departures in the current results from the predictions of the tonal hierarchy (Chapters 4–5). Rather than clearly emphasizing the tonic triad, the pattern of scale-degree expectancies appears more similar to the scale-



degree probabilities observed in several surveys of music. Consistent with the hypothesis that probe-tone methods are sensitive to closure, a survey of *phrase-final* scale degrees in melodies closely mimics the tonal hierarchy.

In order to establish this link more convincingly, a second experiment was conducted to reproduce the important elements of the probe-tone design using the new method. In this case, the tonal hierarchy is clearly reproduced.

The conclusion reached here is that there are at least two distinct schemata in melodic expectancy: continuation, or the expectation the melody will continue, and closure, or the expectation that the melody is ending. This calls into question the global stability of the tonal hierarchy, situating it instead as a model of stability at points of closure.<sup>1</sup>

There are many avenues of further research that await exploration. A number of assumptions were made in the design of the studies presented here that need verification, and there are many variations on the basic experimental design that could further elucidate the mental representation of melody.

One result of this research is to provide a more complete explanatory framework for understanding tonal expectancy. It remains an open question, however, whether the description provided here will stand, or if more expectancy schemata are yet to be found.

---

<sup>1</sup> Brown, Butler, and Jones (1994) theorized that probe tones could measure both continuation and closure, at least with regard to harmony. Closure was defined as harmonic motion to the tonic, in which the prior context is perceived in a non-tonic harmony, and the probe tone is perceived as the tonic resolution. Continuation was defined as the perception that the probe tone is in the same harmony as the prior context.

## CHAPTER 2

### THEORIES OF MELODIC EXPECTANCY

The psychological study of melody perception has a history stretching back over a century. Where nineteenth-century theorists regarded melody primarily from the perspective of “unity” and “coherence,” many late twentieth-century theorists regard melody primarily from the perspective of realized or thwarted expectations. Many of the same questions continue to be topics of concern, however. Whether a melody is said to have “unity” and “coherence,” or to avoid “violations of expectancy,” by any name it sounds just as sweet.

In recent decades, two theories of melodic perception have captured the attention of psychologists. One of these is the tonal hierarchy model, which proposes a static framework of tonal expectancy (Krumhansl, 1990). The more recent is the Implication-Realization theory, which posits, among other things, that melodic contours can be reduced to a fixed set of perceptual archetypes (Narmour, 1990, 1992).

There are obviously many more factors to be considered, and vivid examples have been constructed that demonstrate the effects of rhythm, meter (Francés, 1958/1988), and

auditory streaming (Bregman, 1990) on the perception of melody. Quantitative models exist only for the dimensions of contour and tonality, however, and the present study will be limited to those.

Both the tonal hierarchy and Implication-Realization models have already been subjected to numerous critiques. In addition to those, it will be argued here that the details of the tonal hierarchy have been strongly influenced by task demands.

In order to frame the purpose of the present study, the histories and critiques of both theories and their predecessors will be explored in greater detail below.

### The Grand Illusion: Harmonic Influences on Melody

Modern European music theory is based largely on harmonic analysis. Correspondingly, the earliest empirical studies of music psychology were concerned with the perception of simultaneous tones (Helmholtz, 1877/1948; Stumpf & Meyer, 1898). Following in the tradition of the Greeks and Medievalists, the earliest studies of the perception of melodic intervals naturally attempted to tie melodic structure to harmonic structure. Although by 1898 Carl Stumpf and Max Meyer had demonstrated that ratios have only a passing relation to harmonic intervals, Meyer himself continued to insist on the importance of interval ratios.

Whereas Max Meyer was convinced that harmonic ratios were predisposed to move in the direction of numbers that were powers of 2 (M. Meyer, 1901), in 1910 William

Van Dyke Bingham argued for a less acoustic description. Rather than looking for musical structure at the level of the harmony, he was searching for an explanation that supported an

esthetic unity or wholeness, such as distinguishes a definite melodic phrase when contrasted with a mere fragment of melody, or which characterizes even more clearly a complete melody that is brought into comparison with any portion of itself (Bingham, 1910, p. 20).

Bingham was attempting to identify the perceptual forces that hold a series of tones connected into an apparent whole. Perhaps because of the methodologies he borrowed from the motor action studies of R. H. Stetson, or because of the famous empathic (i.e., embodied) aesthetics of Lipps (whom he cited numerous times), Bingham framed his work as an investigation of a “motor theory of melody.” It seems reasonable to reinterpret this in modern terms as a theory of expectation, focusing on the effects of expectancies on attention and physiological correlates.

Of all the factors influencing aesthetic unity, the one Bingham dwelt on at greatest length was melodic trend. In several of his studies, listeners were presented with an interval and asked, “Does this melody end?” (Permitted responses included affirmative, doubtful, or negative.)<sup>2</sup>

Like Meyer, as well as Lipps, Bingham found that his experimental results were consistent with a preference to end on the side of an interval ratio that was a power of 2.

---

<sup>2</sup> It is a stretch of meaning to call a single melodic interval a “melody,” and a questionable step to apply studies of single intervals to melodies in general. Not all of Bingham’s stimuli consisted of two notes, however.

That is, if the (simplified) frequency ratio of two notes in a melodic interval could be approximated by integers, and if one of the numbers in the ratio could be expressed as a power of 2, then listeners preferred the ordering ending on the note represented by the power of 2.<sup>3</sup>

There were surprising irregularities in responses to Bingham's experiments, however, and he was led to hypothesize that inconsistencies actually resulted from ambiguities in the tonal contexts of the stimuli. In a follow-up experiment, a key was established before each interval. After that, a consistent preference emerged for intervals that ended on the tonic, or members of the tonic triad. For those preferred intervals, the tonic triad members were always powers of two within the interval ratio, which nicely explained Lipps' "law of the number 2."

In addition to melodic trend, Bingham identified a number of secondary factors that influenced melodic coherence, such as the preference for small intervals, and the 'law of the return,' a preference to return to the previous pitch.

The greatest influence of von Bingham's work appears to have been the final rejection of the theories of Lipps and Meyer, as evidenced in the later conversion of researchers such as Farnsworth (1926). It is important to recognize, however, that

---

<sup>3</sup> For instance, the interval C4–G4 is a perfect fifth. The equal-tempered frequency ratio of this interval is 261.63:392.00, which is roughly the same as 2:3 ( $0.667423 \approx 0.666\dots$ ). Of the two halves of the ratio, only the "2" can be expressed as a power of two. Therefore the preferred ordering is 3:2 (a falling fifth), or G4–C4, rather than 2:3 (a rising fifth).

harmony does influence the scale structure of melodies. Huron (1994), for instance, has demonstrated that the Western major and minor scales are the two maximally self-consonant collections of 7 notes from the 12-note chromatic set.

Later in the century, Francés (1958/1988) also recognized a perceptual expectation to have melodies end on the tonic chord. In a vein similar to Bingham's motor theory, he cited Teplov as having identified the phenomenon that melodies ending off the tonic triad are felt "emotionally as a tension requiring a completion — not on a logical level, but on a *sensory level*" (Teplov, 1966, p. 91).

In 1979, Krumhansl and Shepard began a new wave of perception research into melodic trend. They moved from Bingham's 3-point scale to a 7-point scale, and shifted the object of attention from intervals to scale degrees. Instead of asking how well an interval ended, the new focus was on how well the last note completed the pattern. When the pattern was a major scale, their results mirrored those of Bingham, showing a preference for both small intervals and the tonic triad. When the context pattern was changed to a 3-chord cadence played with "Shepard's" tones (Krumhansl & Kessler, 1982), the interval-size effect disappeared, but the subdominant degree gained surprising strength.

A new wave of melodic expectancy research started after Narmour published a pair of books theorizing about the structural nature of melodies (Narmour, 1990, 1992).<sup>4</sup>

---

<sup>4</sup> The term "melodic expectancy" had actually been introduced by Carlsen and others years earlier (Carlsen, 1981; Carlsen, Divenyi, & Taylor, 1970), but without a substantive theoretical framework those studies have been remembered largely for their methods rather than their findings.

Rather than focusing on precise scale degrees or intervals, Narmour's theory instead deals with the general contours of melodies. Relative direction, relative interval size, and their interactions are the primary factors in this model.

Working together, Narmour, Krumhansl, and Schellenberg codified the principles of this theory — the “Implication-Realization” (“I-R”) model — into five testable hypotheses (Krumhansl, 1995; Schellenberg, 1996). In tests of the efficacy of these hypotheses, all five were found to be significant predictors of listeners' responses, after an additional predictor was added for tonal strength. These results were later replicated using other methods (Cuddy & Lunney, 1995; Thompson, Cuddy, & Plaus, 1997).

Together, the I-R and tonal hierarchy models form the basis of most quantitative research into melodic expectancy. Because they are used so extensively in the studies presented here, a more extensive review and evaluation of these two theories is given below. critique

### An Overview and Critique of the Tonal Hierarchy Theory

In 1979, Carol Krumhansl and Roger Shepard published a landmark study of music perception. In it they presented a method for measuring how well each of the 12 chromatic pitch classes fits with a given musical context, using “probe tones.” The resulting graphs showed tantalizing glimpses of diatonic structure, and preferences for particular pitches within the diatonic set.

Consistent with prior research (e.g., Dowling, 1971), the strongest expectancy ratings were given to diatonic (in-key) scale degrees. Two more levels of structure were

also observed: the tonic chord members had higher ratings than the other diatonic degrees, and the tonic had the highest of all. In a follow-up article, Krumhansl and Kessler (1982) defined canonical forms for the major and minor preference ratings based on a small set of chords and chord progressions. The characteristic distribution of scale degree ratings for tonal stimuli was dubbed the “key profile” (Krumhansl & Kessler, 1982), shown earlier in Figure 1.1 (p. 3).

The key profile is sometimes cited as evidence for a “tonal hierarchy” because of its characteristic appearance. The tonic is the highest member of the hierarchy, and the highest rated in the key profile, followed by the third and fifth, and then the other diatonic scale degrees (Krumhansl, 1990). A fourth level of the hierarchy — not pictured in Figure 1.1, because the present study is only concerned with diatonic expectancies — consists of the non-diatonic scale degrees.

Krumhansl’s work provided a new entrée into the systematic study of tonality in musical structure and perception. Prior to that, tonality had been defined in terms of a simple distinction between in-key and out-of-key notes: tonal stimuli were those in which all pitch classes could fit within a single key. Although the key profiles did not give much more explanation about the structure or reason for tonality, they did provide more detailed descriptions of what something tonal might look like.

There is an important distinction to be drawn between “profiles” and “hierarchies.” The key profiles are a matter of fact, the objective result of the particular stimuli and methods used by Krumhansl and Kessler. The tonal hierarchy is a particular theory about the structure of the key profiles in which categorical distinctions are made between levels



of the scale degrees. Although these terms can often be used interchangeably, the term “key profile” is more theory-neutral. Furthermore, profiles other than the key profiles are possible. When stimuli other than those of Krumhansl and Kessler are considered, the results will be referred to as “tone profiles.”

There have been several objections posed to both the probe-tone method and the key profiles, the majority of them expounded by David Butler and collaborators (Butler, 1982, 1989a, 1992; Butler & Brown, 1984, 1994). Among their concerns are the static nature of the key profile, the unaccounted influence of ordering through time, possible confounds with the structure of the test stimuli, a confusion between chords and keys, and inadequate theoretical explanation for the key profiles. An argument that simmered between the Butler and Krumhansl camps for decades peaked in a series of papers and responses between the two which together summarize the major features of the argument (Butler, 1989a, 1989b; Krumhansl, 1989). For that reason, Krumhansl and Butler are represented in the following discussion as the primary figures in the debate.

### The Static Key Profile

The first of Butler’s objections to the key profile is that it is static. To Krumhansl this is a desirable feature, because the key profile is the foundation on which the rest of musical structure is built. In the hierarchy, the tonic is the most stable note of any key, followed by the dominant, mediant, the other diatonic tones, and then non-diatonic tones.

The dynamics of melodic motion from one note to the next are determined by the movement toward or from stability in the hierarchy. The hierarchy of tones is matched by a hierarchy of chords, which sets up its own levels of stability of instability.

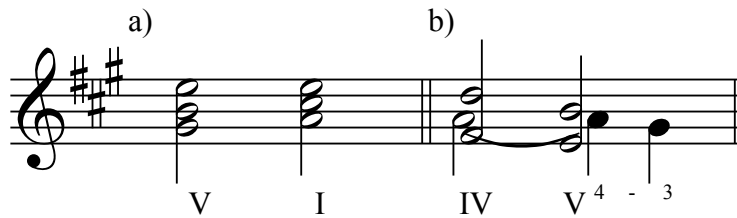


Figure 2.1. Two chord progressions in the key of A major. In (a) the tonic is stable when progressing from *V* to *I*, but in (b) the tonic is unstable when suspended from *IV* to *V*.

The static conception of tonality is untenable to Butler, however. This definition of tonality denies the basic phenomenon that a stable note in one (within-key) harmony can be unstable in another (within-key) harmony. The motion from a *V* chord to *I* (example *a* in Figure 2.1), is often described as a kind of release of tension, because it moves from non-tonic to the stable tonic. In the transition from *IV* to *V*, however, the tonic itself becomes unstable if it is suspended into the *V* (example *b* in Figure 2.1). But the chord progression *I-IV-V-I* is never thought to leave the “home” key, and Krumhansl herself has used the *IV-V-I* progression to “unambiguously establish” a key. What, then, is happening to tonality in the transition from *IV* to *V*? This question will be dealt with at greater length later on, but for now it remains a curiosity of the key profile as it has been traditionally defined.

### Ordering Through Time

The second problem with the static nature of the key profile, according to Butler, is its silence on the effects of the ordering of pitches through time. “Krumhansl and Shepard look for tonality-imparting information in the tones, not in the relationships among them,” he points out (Brown & Butler, 1981).

The hallmark of the key profile is its stasis: once a tonality is invoked for listeners, testing will be able to “recover” the outlines of the key profile from their responses. However, Butler has highlighted a number of experiments that demonstrate the recovery of the key profile is not always so robust. Cuddy and Badertscher (1987) tested children and adults using an arpeggiated major triad to invoke a key, and found a reasonably good fit to the major key profile. Brown, Butler and Jones (1994) replicated this finding, and then changed the ordering of the arpeggiation. In the revised arpeggiation the clear orientation of the tone profile toward the tonic disappeared, and resulted in a significantly lower correlation to the original arpeggiation. In this same study, a reordering of the arpeggiated diminished triad used by Cuddy and Badertscher also resulted in a significantly different tone profile. In another study it was found that reordering the notes in two dyads to create a tritone will significantly improve listeners’ ability to hear an unambiguous statement of key (Butler, 1982). And the same set of 3 to 10 notes can alternately imply one key or another, or a large number of keys, depending on their order (Brown, 1988).

Krumhansl disputes the idea that the key profiles are insensitive to ordering effects. For instance, there is an asymmetry in the similarity between two pitches, depending on

whether the less stable pitch is first or last (Krumhansl & Shepard, 1979). This is an important phenomenon because among music theorists the motion from unstable to tonic is “unanimously regarded... as important in shaping the flow of music in time.” A similar ordering effect has been observed between chords, where two chords are rated as more similar if the second chord is higher in the harmonic hierarchy (Bharucha & Krumhansl, 1983; Krumhansl, Bharucha, & Castellano, 1982; Krumhansl, Bharucha, & Kessler, 1982).

A number of other studies are cited as examples of ordering effects in music perception, including memory for tones (Bartlett & Dowling, 1988; Dowling & Bartlett, 1981; Krumhansl, 1979), and memory for chords (Bharucha & Krumhansl, 1983; Krumhansl, Bharucha, & Castellano, 1982).

Butler points out that none of the memory experiments systematically varied time-orders of tones. Rather, they simply substituted one tone for another in the comparison stimulus. In all of these experiments it was found that a change from an unstable (read: non-diatonic) element in the standard stimulus to a stable one was less likely to be noticed than a change from a stable element to an unstable one. This is an example of an “ordering effect,” according to Krumhansl. More likely, as Bartlett and Dowling (1988) have argued, it is because a standard composed entirely of diatonic elements generates strong expectations for the diatonic set. A standard including a non-diatonic element broadens the listener’s expectations, and makes changes in the comparison less noticeable.

The last problem with the “temporal ordering” memory studies is that they only test the lowest two levels of the tonal hierarchy, namely the diatonic/chromatic distinction. As Butler points out, this is a confusion of “tonality” with “diatonicism.” If there is a demonstration of the effect of tonal hierarchy here, it is merely that there are two hierarchical levels: in-key, and out-of-key. Without the additional hierarchical level of tonic-dominant-mediant, the tonal hierarchy would simply be a theory of what notes are in the key.

Other researchers have managed to incorporate timing elements by expanding the scope of the original key profile model. Huron and Parncutt (1993) hypothesized that listeners might have a “window” of perception (something like the auditory sensory memory) that sums note durations, weighted by recency, and matches key profiles from moment to moment. This resulted in better performance than the original model (an algorithm defined in Krumhansl, 1990) for predicting listeners’ responses to the stimuli of the Butler (1982) dyads and the Brown (1988) ambiguous melodies. Performance on Brown’s less-ambiguous melodies was not improved, which led the authors to conclude that tonal phenomena with temporal-dependent factors — what Brown (1988) called “functional” tonality — could not be adequately modeled using the key profile approach alone.

#### Effects of Stimulus Structure

Another of Butler’s concerns is that the probe-tone task might not reflect abstracted, generalized knowledge about tonal structure, but rather a surface response to the

immediate context. The ratings provided by the listeners have a striking similarity to the duration each pitch class was sounded within the stimuli. In addition, the direction, contour, repetitions, implied harmonies, and serial position of elements in the stimuli might all influence responses. Butler has pointed out that in one of the original stimuli, the ascending major melodic scale pattern (see Figure 2.2), the tonic is sounded twice. He hypothesizes listeners might choose the tonic as a good completion note for that reason alone.



Figure 2.2. The ascending major melodic scale pattern (from Butler, 1989a).<sup>5</sup>

---

<sup>5</sup> This notation is slightly different from the stimuli used by Krumhansl and Shepard (1979). Their probe tone was played an octave lower than notated in Butler's example.

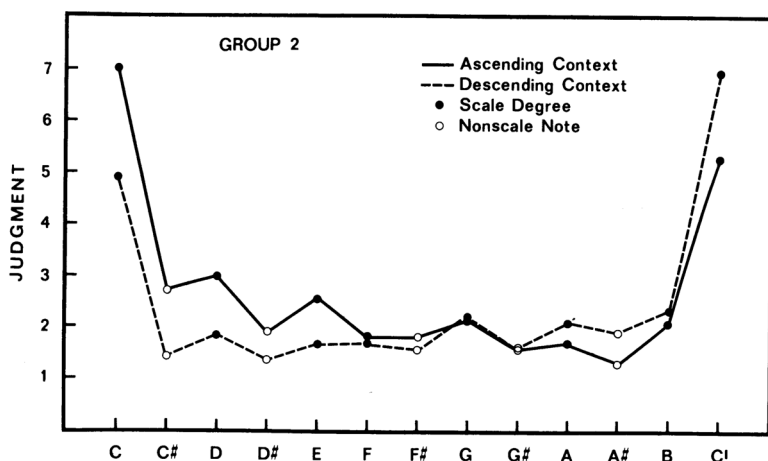


Figure 2.3. The tone profile for Krumhansl & Shepard (1979) major melodic scale patterns, for subjects with moderate musical training (Group 2).

Krumhansl has claimed that the duration counts for the ascending major scale could not account for the major-key profile. But all melodic contexts exhibit pitch proximity effects, and no scalar context ever produced a canonical key profile. The tone profile predicted by the duration *was* observed for some subjects in the original Krumhansl and Shepard (1979) experiments. Subjects in that study were divided into three groups based on musical background. Group 2 responses (3.3 years performing experience) produced the pattern predicted by Butler, namely strong ratings for the tonic, and flat ratings for the other degrees (see Figure 2.3).

To minimize the influence of pitch proximity, the key profiles were derived using chords rather than melodies, and stimuli were played using “Shepard’s tones,” a type of pitched sound having octave ambiguity (Shepard, 1964). Both of these minimized the

potential for consistent confounds based on contour influences. When melodic contexts are used that have clear contours, the resulting profiles can vary widely (Brown et al., 1994; Cuddy & Badertscher, 1987).

There is evidence that key profiles represent abstracted mental representations, according to Krumhansl. This can be found in the correlations among probe-tone profiles and duration counts. Whereas the inter-tone-profile correlations are high in the Krumhansl and Kessler (1982) data ( $r = 0.9$ ), the inter-duration-count correlations for the stimuli are lower ( $r = 0.75$ ). If responses were determined by the duration counts of the stimuli, then the inter-profile correlations should not have been higher.

Krumhansl and Kessler selected the highest inter-profile correlations post hoc, however, so it is entirely possible the lower inter-duration correlations are lower completely by chance. In addition to omission of the scalar stimuli, Butler noted that the diminished and dominant-seventh chords were also not included in the key profile average. That raises the question, do diminished and dominant-seventh chords fail to establish a key, or is it that probe tones are not really measuring key?

Thomson (2001) noted that although the 3-chord cadence was intended to establish an unambiguous key center, it was not necessarily successful in doing so. A *IV-V-I* cadence in C major is identical to a *I-V/V-V* progression in F major, both of which are extremely typical in nineteenth-century practice. This is exactly the problem Bingham faced but in another guise: in one hearing the cadence may imply C, but in another it might imply F. If that were the situation, the resulting preferences could be a mix of a C-major triad and an F-major triad.



In one study, Povel (1996) asked listeners with performance experience but no formal musical training to first listen to a 4-chord progression followed by one of the 12 chromatic tones, and then play their preferred completion on a keyboard. Povel then separated his participants into three groups based on their tendency to prefer the tonic triad. The largest group (45%) showed a clear preference for only the tonic, dominant, and mediant scale degrees, in descending order. The second group (29%) always preferred an upward continuation by fourth, regardless of the tonal relation of the prompt tone to the chord sequence. The smallest group (23%) showed a general preference for diatonic completions, but were otherwise random. These results suggest there are some listeners who prefer to hear *V-I* completions regardless of context, and other who prefer tonic-chord completions in the established key. This supports Thomson's contention that multiple tonal contexts can be read into a single "clearly established" tonal center, and suggests a reason for the complex zig-zags of the key profiles.

Krumhansl cites several studies that replicate the Krumhansl and Kessler work, and offer evidence that short-term context effects fail to predict experimental data. These studies do not clearly counter the charge, however. Cuddy, Cohen, and Miller (1979) measured the effect of diatonic versus non-diatonic changes, as well as the effect for closure on tonic, but neither of those establishes the multi-tiered tonic-chord/diatonic/non-diatonic hierarchy of Krumhansl and Kessler. Krumhansl's (1979) memory task similarly only considered the differentiation between diatonic and non-diatonic tones, the bottom two levels of the hierarchy.

Palmer and Krumhansl (Palmer & Krumhansl, 1987a; 1987b) and Schmuckler (1989) all produced results that correlated significantly with the key profiles. However, it is important to note that a good correlation is possible even if there is almost no hierarchical structure to the responses. A more detailed analysis would be necessary in order to accept Krumhansl's claim that these studies substantiate the key profiles.

Janata and Reisberg (1988) used a reaction-time study to measure expectancy sensitivities to scale degrees. Rather than analyzing the ratings subjects assigned to probe tones, they studied the amount of time it took subjects to decide on a rating. The reaction times replicated the findings of Krumhansl and Kessler for both isolated chords and scalar patterns. This demonstrates that the vagueness of the rating task is not responsible for the shape of the key profile.

The obvious retort to these studies, made earlier, is that there are many experiments that have indeed shown the influence of short-term, stimulus-specific context effects (Brown, 1988; Brown & Butler, 1981; Brown et al., 1994; Butler, 1982). In turn, it has not been sufficiently demonstrated that the key profiles are independent of the particular stimuli used to create them.

### Equating Chord With Key

As mentioned earlier, the key profile is static by definition. One consequence of this is there are only two ways to describe harmonic change using the key profiles: either as motion to and from unstable harmonies, or as jumps through foreign key areas.

In their 1982 paper, Krumhansl and Kessler opted to present within-key harmonic motion as if it were moving through foreign keys. In Figure 2.4, a major diatonic chord sequence in C major is shown moving through tonal space. Notice how the motion through chords 6, 7, and 8 (*vi—ii—V*) have the listener leaping from C major to somewhere in the middle of A minor, D minor, and F major, and then back again. Of course, given how meager our current understanding of harmony perception is, this may indeed be a plausible hypothesis: motion from tonic to subdominant function, for instance, could actually be heard as a fleeting modulation to the subdominant key.<sup>6</sup>

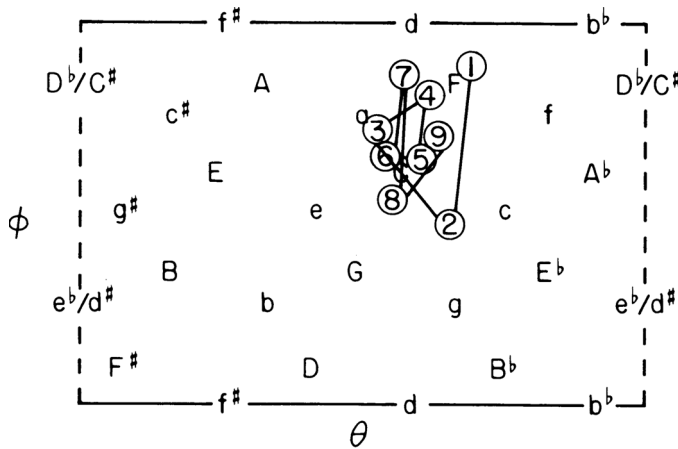


Figure 2.4. The path of a diatonic harmonic progression (*IV—V—vi—IV—I—vi—ii—V—I*) through foreign key space (from Krumhansl & Kessler, 1982).

Butler has proposed that listeners are not actually moving through foreign key space, but rather misattributing tonic qualities to whatever chord a sequence stops at. For

<sup>6</sup> This idea is more similar to Rameau's (1722/1971) original conception of modulation than to the modern definition.

example, chord 7 in Figure 2.4 is clearly serving a subdominant function in a C-major chord progression. If the sequence were to stop at chord 7, however, it would be trivial to reinterpret the final three chords, *I-vi-ii*, as *VII-v-i* in D minor, which is a reasonably satisfying minor cadence. This could explain why chord 7 is placed near D minor in the figure. It is also a more appealing notion than positing that harmonic motion must always be analogous to modulation.

According to Krumhansl, key profiles can be used to separate the effects of local tonicization from the effects of the key. The correlation calculated between the ostensibly prevailing key profile and an individual chord is subtracted from the correlation between the key profile and the listener's probe-tone profile at that chord. These two levels — the chord and perceived key — are hypothesized to exist simultaneously in the sense of Schenker's (1935/1979) levels.

Contrary to the tonal hierarchy model, however, there is evidence that separate tonal contexts do exist for each harmony (Holleran, Jones, & Butler, 1995; Palmer & Holleran, 1994; Trainor & Trehub, 1994). In each of these studies, it was found that confusion errors were more likely for tones that were within-harmony, rather than merely within-key. According to Krumhansl (1990), the tonal hierarchy has levels for non-diatonic tones, diatonic tones, and finally tonic-chord tones. But the findings of Holleran, et al., apparently necessitate some amendment of this hierarchy to reflect dynamic changes in harmony.

## Problems Explaining the Key Profiles

What are the perceptual origins of the key profiles? According to Krumhansl (1990), the distribution of notes in music leads us to perceive more common tones, such as the tonic, as more stable. The key profile is in effect learned from the frequency of scale degrees in music. Butler has called attention to the therefore surprising lack of correspondence between key profiles and the frequency counts cited by Krumhansl.

Hughes' (1977) analysis of the first *Moments Musicaux* by Schubert was cited as a paradigmatic example of the match between key profiles and tone distributions (Krumhansl, 1987).<sup>7</sup> The duration count of the pitch classes is an “almost perfect” correspondence to the key profile for G major. In addition, Hughes explicitly identified an “orientation” toward G major in the piece (even though it is written in C major). Butler points out, however, that the piece really does spend the first few measures in C major, and then moves through C minor, D major, E-flat major, E minor, G minor, and A minor. What do we learn from knowing that the “orientation” of the piece is toward G? A key profile analysis would claim the orientation of the first 8 measures is toward G major, even though it is clearly in C (with hints of C minor), simply because of the predominance of the note G. Butler notes that the most common tone for both the first and second *Moments Musicaux* is the dominant, and suggests this is because the most common chords are *I* and *V*, and the only common tone between them is the dominant.

---

<sup>7</sup> This citation courtesy of Butler (1989b).

Krumhansl's wider point is that the correspondence between the key profiles and the distribution of notes in musical practice is strong. Two prior surveys substantiate this claim: those of Youngblood (1958) and Knopoff and Hutchinson (1983). Those two included some 25,000 tones taken from the melodies of Schubert, Mendelsson, Schumann, Mozart, Strauss, and Hasse vocal works (with some overlap). The correlations with the key profiles are high, upwards of 0.86.

What does it mean to have a correlation greater than 0.8? Consider that of the 14 tone frequency tabulations in these two publications, only two rank the tonic as the most-common tone, only eight rank the tonic triad members as the three most-common tones, whereas 11 rank the dominant as the most common. This is a problem if the key profiles are learned from statistical properties of real music. It does put the dominant orientation of the *Moments Musicaux* in perspective, however.

The reason for making these associations between musical statistics and perceptual studies is that Krumhansl believes learning the frequency of musical events is an important step toward becoming aware of principles of musical organization such as cadences, temporal ordering, implied harmony, meter, and rhythmic stress. All of these may work together to create a sense of tonality.

This may be true, but it is not necessarily the case that Krumhansl's measures of tonal and harmonic stability capture that acquired statistical knowledge. There is apparently a disconnect between the frequencies of events and the ratings given them by listeners.

### An Overview and Critique of the Implication-Realization Model

The Implication-Realization model is more recent than the key profiles, but its pedigree stretches back nearly a half-century. The importance of the realization (or subversion) of expectation in music was introduced by Leonard Meyer, Narmour's mentor, in his book *Emotion and Meaning in Music* (1956). One of the arguments presented there is that emotion and meaning are created through expectation, that "one musical event (be it a tone, a phrase, or a whole section) has meaning because it points to and makes us expect another musical event" (p. 35).

In addition to this general conceptual framework, Meyer provided Narmour with an important dichotomy: he believed that style is learned, but music perception is also determined by Gestalts which are innate and therefore universal. Although Meyer makes no decisive claims about the mechanisms of either learning or Gestalts, Narmour applies a cognitive interpretation, mapping Gestalts onto "bottom-up" processes, and style onto "top-down" cognition.

Narmour first became (in)famous in music theory circles for his heretical text, *Beyond Schenkerism* (Narmour, 1977). According to the book, one of the faults of Schenkerian theory is its lack of any solid foundation in real-world meaning. Despite this criticism, Schenkerian thinking remains popular because of its facility for imputing structural meaning to the smallest details of contrapuntal, and hence melodic, organization. Narmour's I-R theory was intended to provide an alternate means of conducting detailed musical analysis, one with a firm footing in the empirical world of cognitive science.

The full Implication-Realization theory has a great deal of complexity, incorporating tonal, metric, harmonic, and style components, to name a few. Empirical studies have been limited to contour, however, which is the backbone of the theory.

The I-R theory posits the existence of numerous archetypes which take the form of common one- and two-interval melodic patterns. These archetypes are explained by Narmour as products of primitive processes (Gestalt laws working at a pre-attentional level) and stylistic convention. There are both implicative and retrospective archetypes, but only the prospective archetypes will be considered here.

Of the basic archetypes, only three have both implication and realization components, namely process, reversal, and registral return. Both process and reversal each have two “derived” archetypes that realize expectancies along only one of the two dimensions of “interval” and “register.” The resulting seven prospective archetypes are shown in Table 2.1.



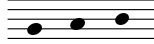


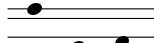
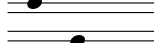
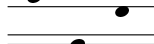
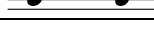
Archetype	Implicative	Realized		Diagram
	Interval	Interval	Direction	
process	Small	Small	Same	
intervallic process	Small	Small	Opposite	
registral process	Small	Larger	Same	
reversal	Large	Smaller	Opposite	
intervallic reversal	Large	Smaller	Same	
registral reversal	Large	Larger	Opposite	
registral return	Any	Similar	Opposite	

Table 2.1. The Seven Prospective 3-Note Archetypes of Narmour's Implication-Realization Model

The precise definitions of these archetypes have been provided by Narmour (1990). According to Narmour's theory, archetypes function as prototypes, and some variants are theorized to fulfill expectancies better than others.

There are three primary critiques of the I-R model. Because Narmour bases his theory on Gestalt principles, it is subject to all the criticisms normally leveled at them, as well as a few more because of the particulars of Narmour's interpretation. Furthermore, the model has a considerable amount of redundancy, both in its initial form and in the model first presented by Krumhansl (1995) and Schellenberg (1996). Finally, the major factors of the theory, interval size and direction, have been shown to be roughly equivalent to the more general principles of inertia and regression to the mean.

## The Gestalt principles

The ‘Implication-Realization’ (I-R) theory is based on three purportedly innate Gestalt principles: similarity, proximity, and common fate or ‘common direction.’ (A fourth principle of ‘reversal’ is proposed as a hitherto unidentified innate expectancy.) Narmour proposes these principles can operate on any dimension or ‘parameter scale’ imaginable: melody (especially interval and direction), duration, register, dynamics, texture, timbre, meter, tessitura (Narmour, 1990, p. 13), harmony (especially chord inversion, soprano-bass relations, and dissonance), tempo, voice-leading, common-toneness, spacing, performance attack (pp. 288–289), virtual register (i.e., degree of melodic closure, p. 363), to name those explicitly mentioned. Furthermore, “every piece [of music] involves implication in some significant sense in every note-to-note relation of every parameter on every level” (p. 13). Proof of the reality of these scales and principles is not a priority, however, because “such Gestalt laws, in fact, need no defending in the cognitive psychology of music. They form a major part of the canon of observed phenomena for music perception” (p. 63). The cited evidence proving the existence of bottom-up Gestalt mechanisms is an article by Pomerantz (1981) on visual perception which identifies similarity, proximity, and common fate as likely data-driven mechanisms.

Inspired by cognitive psychology, Narmour posits a two-tier system of perceptual processing, one top-down, and the other bottom-up. In particular, “the top-down one is

flexible and variable but controlled; the bottom-up one is rigid, reflexive, and automatic — a computational, syntactic input system” (Narmour, p. 54). The bottom-up system is a kind of General Computer of expectancy.

Although fixed or ‘impenetrable’ low-level processes have been proposed as a model of brain function (e.g., Fodor, 1983), their existence is hardly a given. Even Pomerantz, Narmour’s source of evidence for innate Gestalts, thought that consciousness “instructs preattention how to behave” (1981, p. 151). But according to Narmour, the expectancies generated by basic melodic forms are fixed, and exceptions must be culturally learned, and are therefore top-down. Following in the footsteps of Meyer (1956), this assumption conveniently explains how sensations of “surprise” can still exist even in the face of strong veridical expectancies.

The I-R principles are derived from Gestalt mechanisms by appeals to both intuitive and formal logic, and both methods have their failings. The basic I-R melodic structure is ‘process,’ a small interval followed by another interval of similar size in the same direction. The importance of process is explained by the parameters of pitch and interval using all three Gestalts. The intuitive appeal works like this: small intervals consist of pitches that are *proximate* in pitch, which generates an expectation for a continuation of *similar* interval size; and the direction of the interval creates an expectation that the next note will continue in the *common direction*.

In the first place, the logic here is somewhat troubled. Secondly, no other account of Gestalt psychology would consider the apparent motion of a single object from one place to another an example of common direction (or common fate) — rather this would be

better termed “good continuation.” But Narmour has rejected principles of “good continuation,” “good figure,” and “best organization” as top-down, interpretive, and unreliable (p. 63).

The formal argument for the establishment of basic melodic structures uses a pseudo-propositional-logic notation. Continuation is implied by a small interval because  $a + b \rightarrow c$ , where  $a$ ,  $b$ , and  $c$  are pitches; and repetition is implied because  $a + a \rightarrow a$ . Or rather, continuation is implied because  $A(a + b) \rightarrow A(b + c)$ , where  $A$  is a small interval. The usefulness of these derivations has been questioned by Thompson (1996, p. 142), who suggests “further consideration of the universal hypotheses” is necessary.

In short, Narmour’s invocation of “established” psychological principles is unsatisfying from a scientific viewpoint. But even though the justifications for hypotheses are repeatedly sought via cognitive principles, Narmour often seems to be appealing to the musical sensibilities of theorists. It may be more useful to cast aside the cognitive trappings of the theory, and instead view it as a systematic codification of the author’s musical intuitions.

### Redundancy

The first testable form of the I-R model was developed by Krumhansl (1995) and Schellenberg (1996). The two hypothesized dimensions of contour are “interval” and “register,” which measure the size and relative direction of intervals, respectively. The factors constructed from these dimensions were called intervallic difference and registral direction. Intervallic difference encodes Narmour’s dictums that small intervals beget

small intervals in the same direction, and large intervals are followed by small intervals in the opposite direction. The registral direction parameter simply codes the rule that small intervals are followed in the same direction, and large intervals are followed in the opposite direction. If a problem with these definitions seems to be lurking, it will be made explicit later.

The registral return parameter was encoded directly from Narmour's archetype of the same name, encapsulating the intuition that a melodic interval is likely to be followed by a return to the first pitch of the interval, regardless of the interval size. In the I-R theory, a large interval followed by a reversal of direction and interval size creates "closure," a marker for perceptual phrase boundaries, so a pitch closure factor was included in the model. Finally, it was noted that pitch proximity (a tendency toward small intervals) is implicit in many of the archetypes, so a pitch proximity parameter was added.

The Implication-Realization (I-R) model was first tested by Krumhansl (1995) and Schellenberg (1996), who found support for the principles they tested, except that unisons (repeated notes) were apparently less expected than predicted. These same five parameters have been tested and confirmed by others using different methods (Cuddy & Lunney, 1995; Thompson et al., 1997).

In his analyses, Schellenberg (1996) noticed considerable redundancy among the intervallic difference, proximity, and closure predictors. In a revised model, this was corrected by discarding the intervallic difference and closure hypotheses, and recoding proximity. "Closure," after all, is basically a redundant measure of registral direction

(reversal is expected after large intervals); and intervallic difference always predicted intervals would be followed by small intervals, which is redundant with proximity. Another change in the revised model was to recode registral direction so that it only predicted the reversal implication of large intervals, and not the continuation implications of small intervals. This may have been done in order to maximize its correlation with the omitted “closure” predictor. The three-predictor model (plus tonality) was as good as the five-predictor model at explaining listeners’ responses.

In a follow-up study, Schellenberg (1997) used a principle-components analysis to reduce them to two parameters which he named pitch-proximity and pitch-reversal.

As formulated by Schellenberg, pitch-proximity is operationalized as the number of semitones between the current pitch and the previous pitch.

The pitch-reversal predictor is an amalgam of several distinct parameters from the original model, coded with values from  $-1$  to  $2.5$ . According to this principle, listeners expect melodic intervals to be followed by intervals of similar size in the opposite direction ( $1.5$ ), especially for large intervals ( $2.5$ ). To a lesser extent listeners expect large intervals to be followed by intervals of contrasting sizes in the opposite direction ( $1$ ). Large intervals are not expected to be continued in the same direction ( $-1$ ), and there are no predicted expectancies for other interval combinations ( $0$ ).

### Regression to the Mean

The “reversal” rule was proposed by Narmour as a potentially major new contribution to Gestalt psychology. According to his definition, a large change along one parameter implies a reversal along one or more parameters.

According to von Hippel and Huron (2000), the behavior of reversal is already well known as regression toward the mean. They observed that large pitch intervals are more likely to finish at a relative extreme of the melody’s pitch range than are small changes. Assuming a distribution with a central tendency, that means the majority of possible next pitches lie in the opposite direction. Although Narmour claimed that any kind of large change will result in reversal, according to regression to the mean only a change ending toward a tail of the pitch distribution is likely to do so.

As for the tendency for reversals to be small intervals, von Hippel and Huron observed that smaller intervals are more common than large intervals, so a small interval *is* more likely to follow a large interval, or any interval. In a survey of real melodies, they found that there was no evidence of a principle of reversal above and beyond regression to the mean. Most of the Krumhansl and Schellenberg results were also consistent with regression to the mean rather than reversal (von Hippel & Huron, 2000). They concluded there is no need to posit a principle of reversal, since pitch reversals can be adequately explained by basic statistical properties of melodies.

More recently, von Hippel (2002) constructed a set of melodic sequences and manipulated the ending note of each sequence to place it above, below, or on the mean of the sequence’s pitch distribution. Subjects were asked to describe their expectations for

the following note. The results suggested that reversal may in fact be used as a perceptual heuristic, rather than regression to the mean — but only for trained musicians. (Non-musicians demonstrated no significant expectations.) If reversal is learned, however, that undermines its status as an innate process. In addition, von Hippel found evidence that ‘process,’ or the tendency for small steps to be continued in the same direction, also has perceptual validity, at least for musicians.

Both the I-R and tonal hierarchy theories have their critics, but versions of both continue to hold currency in music research. No alternatives to the I-R theory have been shown to perform significantly better, and no successor to the key profiles has been quantified in as appealing a package.

Nevertheless, one of the objectives of the present study is to call elements of these theories into question and to provide alternate accounts. Before discussing any new research, however, it will be important to review the research methods used to measure melodic expectancy, and consider how they might inadvertently redefine the questions they are intended to answer.



## CHAPTER 3

### METHODS OF MEASUREMENT

There has been a kind of Uncertainty Principle in the measurement of pitch and time of melodic expectancy. Even if knowledge of one does not completely preclude the other, the two dimensions are difficult to manipulate independently. The most popular melodic expectancy design, the “probe tone” method, can determine how expected a particular pitch is, but that pitch must be fixed at the end of the stimulus (e.g., Krumhansl & Shepard, 1979). In other designs, expectancy can be measured in the middle of a musical phrase, but the pitch of the observed note cannot be fixed, or a particular event cannot be selected for observation (e.g., Unyk & Carlsen, 1987, experiments 1 and 2, respectively), or measurements cannot be easily assigned to particular events (Eerola, Toiviainen, & Krumhansl, 2002).<sup>8</sup> At the altar of balanced designs, the choice is often made to sacrifice time in favor of pitch.

---

<sup>8</sup> One possible exception is the delayed-recognition test (Cuddy et al., 1979; Krumhansl, 1979; Schmuckler, 1989). Memory errors are influenced by expectancy, and when constructing lures specific pitches can be manipulated. This design has not proved as popular as the probe-tone method, however, perhaps because it is even more time-consuming.

### Previous Methods for Measuring Melodic Expectancy

Some of the earliest designs for measuring melodic expectancy asked trained musicians to indicate their expectations by playing improvised continuations to melodies (Carlsen, 1981; Carlsen et al., 1970). Listeners' errors in transcribing melodies (Unyk & Carlsen, 1987) or imitating them (Mitroudot, 2001) have also been used to study expectancy. For non-musicians, errors in change detection between two melodies have been used as measures (Schmuckler, 1989), and ERP studies of "early P-3" in the auditory cortex have been applied in both attention and expectancy research (Besson & Faieta, 1995).

Most tests of melodic expectancy have been conducted using the "probe tone" method, which was pioneered by Krumhansl and Shepard (1979). In this design, a musical context is presented, followed by a tone, and the listener is asked to rate how "similar" the tone is to the previous tone, or how well it fits the context. During the experiment the listener is usually asked to repeatedly rate the context in relation to all twelve chromatic pitch classes. Compared to most of the other methods, this has the advantage that all possible continuations of the context are measured. Most recently, the rating scale approach has been used in a continuous-response task in which subjects moved a slider up and down while listening to a melody to indicate their surprise from moment to moment (Eerola et al., 2002).

Each of these methods of measuring melodic expectancies has its drawbacks. For some, expectancies can be measured only in terms of the notes the subjects choose to perform or fail to remember. For continuous-response tasks, it can be difficult to determine which notes the ratings should be assigned to.

A larger problem with the probe tone method is that it stops the melody before the listener is asked to respond. This may have the effect of indirectly communicating that the melody has ended. Listeners can then form a retrospective perception of closure when the stimulus ends. In Western music, phrase endings tend to close, or “cadence,” on a stable harmony and scale degree, often the tonic chord.

#### A New Method

A new method was designed that can measure melodic expectancy for specific notes without stopping the melody being studied. In this design, expectancy measures are taken after each note in a continuing melody. This approach has a number of advantages. First, the time pressure of responding after each note is an impediment to engaging in retrospective reinterpretation of the context. The only cues that the melody might end are those the subject might hear in a normal listening context.

The dependent measure in this design is reaction time. Reaction time is a common measure in research on expectation and attention, but has rarely been used in the study of melody. (A notable exception is a study of decision response time in a probe tone task by Janata & Reisberg, 1988.) To be feasible, a task was needed that would require listeners to perceive the pitch of each note before responding, but would be simple enough to be

completed before the onset of the next note. One of the easiest types of coding in the perception of melodies, and perhaps the earliest stage of processing, is contour (Dowling & Bartlett, 1981). As a musical feature, contour also has the advantage that it can be defined in very tangible terms. Determining whether a note moved up, down, or stayed “the same” is a task that is well within the reach of most undergraduate music students.<sup>9</sup>

This method, measuring reaction times to the 3-response categorization task of determining the contour of each note in a melody, was used to estimate expectancies in the situation where subjects expected the melody to continue after each note. As established in prior research (D. E. Meyer, Osman, Irwin, & Yantis, 1988), faster reaction times were interpreted as indications of stronger expectancy. Similarly, slow reaction times were construed to be the result of expectancy violations.

The results will be compared to those obtained from other methods, first to validate the approach, and then to highlight important differences. A second experiment was carried out using the same procedure, but this time using a task designed to mimic the effects of a probe tone design. Some important similarities will be shown between the results of this second experiment and those of probe tone experiments.

To test the new reaction time method, Schellenberg’s two-factor I-R model was selected, for reasons of simplicity. Out of concern generated by the results of von Hippel’s controlled-distribution experiment, however, the process archetype was added

---

<sup>9</sup> As most aural skills instructors can attest, however, even up and down can be difficult for some students to hear.

as a third term in the model. Using Narmour's definition, 'process' was operationalized as an interval smaller than 6 semitones followed in the same direction by an interval at most 3 semitones larger or smaller in size.

As in previous tests of the I-R model, the key profile was included in the model as a predictor of tonality. The other I-R models (Narmour's original archetypes, the five-factor parameter model, and the one-factor statistical model) will be addressed at greater length in Chapter 6.

## CHAPTER 4

### EXPECTANCY FOR CONTINUATION (EXPERIMENT 1)

The purpose of the first experiment was to determine the efficacy of the new experimental design, which measures reaction times in a speeded decision task involving categorization judgments about the pitch of notes. The basic task was to determine whether each note in a melody moved up or down from the previous note or stayed at the same pitch. A listener had to indicate this decision by pressing one of three buttons as soon as possible. The three button options were ‘up,’ ‘down,’ and ‘same.’ The assumption of this design is that a fast reaction time indicates a strong expectancy. Contrapositively, an expectancy violation is assumed to cause a slow reaction time.

The first goal of the experiment was to test whether reaction time measurements to the contour task resemble those of previous melodic expectancy tests. Two categories of predictors were used, namely the Implication-Realization (I-R) model and a tonality model. The I-R model included Schellenberg’s (1997) two-parameter model (pitch-proximity and pitch-reversal) as well as Narmour’s process archetype, for reasons discussed earlier.

Tonality was modeled by the major key profile (Krumhansl & Kessler, 1982). It was decided to test only major-key melodies in order to reduce the number of analyses

and stimuli. The minor-key tonality has several variants and more often includes chromatic alterations in Western music, so the major tonality is a better starting point for the present study.

Previous research has shown that unisons (repeated notes) are not expected events (Krumhansl, 1995), contrary to the predictions of the original I-R model. We would expect this to be reflected in longer reaction times for unisons. A unison covariate was therefore added to the model, coded as a dummy variable.

### Methods

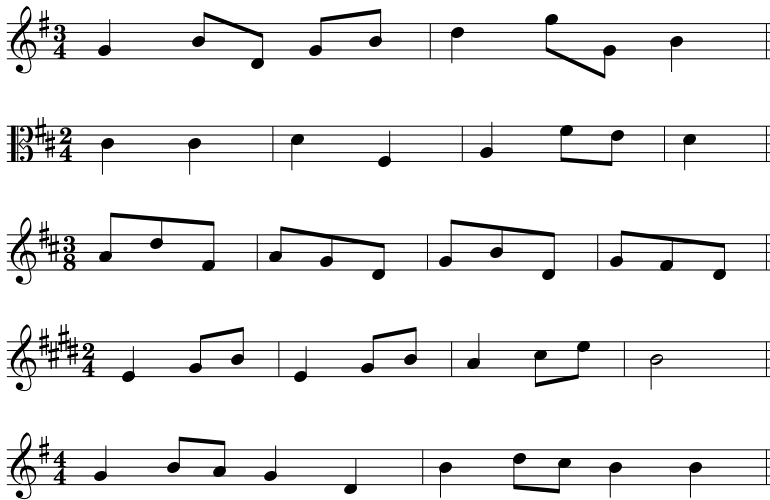
The task of determining whether notes move up or down may seem trivially easy, but in practice it can be difficult even for trained musicians to do quickly and repeatedly. In order to maintain the attention of subjects, the experiment was presented as a game of skill. No prizes or other material incentives were offered, but previous “high scores” were visible to participants. In post-experiment interviews with participants, it became evident that this technique was surprisingly effective at galvanizing their efforts.

### Stimuli

Melodic phrases were sampled from the Essen Folksong Collection (Schaffrath, 1995) using the Humdrum Toolkit (Huron, 1993). The Essen Folksong Collection contains over 6,000 traditional European folksongs, primarily of German origin. Phrase units are coded explicitly in the collection, so no special interpretation was needed. Only the phrases that were in a major key and contained at least 8 notes and no rests were retained. If a phrase contained a note with either a duration that could not be expressed as

multiples of an eighth note, or a duration longer than a half-note, it was excluded. Only those phrases with a leap of at least a perfect fifth (7 semitones) were included, in order to provide an adequate test of the large-interval claims of the I-R model.

Thirty-seven phrases were sampled from the resulting set of melodic phrases. (Some sample phrases are shown in Figure 4.1.) The original keys and ranges were retained for each phrase.<sup>10</sup> The phrases were played at a tempo of 60 eighth notes per minute (i.e., 1000ms per eighth note). The average preferred tempo for many individuals has been estimated at around 100 beats per minute (Fraisse, 1957/1963), but a slower tempo was found to be necessary for the task.



**Figure 4.1.** Five of the 37 melodic phrases from the Essen Folksong Collection used as stimuli in Experiment 1.

---

<sup>10</sup> There were a total of nine keys used among the stimuli, but over three-quarters of the stimuli comprised only four of the keys, namely D, E, F, and G major.



Each melodic phrase was preceded by a four-chord progression (I–IV–V<sup>7</sup>–I) to establish the key of the phrase. Each chord was an eighth-note in duration. The progression was followed by two eighth-notes of silence, and then the melody.

### Equipment

The experiment was conducted on a Linux workstation using software written in Perl and a GUI implemented in Perl/Tk. Measurement accuracy was estimated at around 10ms by graphing reaction times in ascending order and looking for regular discontinuities (Myors, 1998). Melodies were played over headphones in a sound-attenuated booth using the default piano patch of a SoundBlaster AWE card. Responses were entered using a computer keyboard.

### Subjects and Procedure

The subjects were 27 second-year undergraduate music students at the Ohio State University who participated in the experiment for course credit. Presentation order of the 37 melodies was randomized differently for each subject.

Each melodic phrase was presented as a round in the game. Subjects were instructed to determine the contour of each note in the melody and enter their responses on a computer keyboard. Reaction times were measured from the onset of each note. After each response, one point was awarded for each millisecond the reaction time was under 1000 ms.

A pilot study provided evidence that the choice of response hands should not affect the results, even though it was thought that index-finger responses might be significantly

faster than ring-finger responses. The response hand was specified as a between-subjects factor for six right-handed undergraduate students. Subjects were asked to use their index, middle, and ring fingers to enter responses, and to consistently use the same finger to press each key. The right button indicated up, the left indicated down, and the middle button indicated same (unison). The dependent measure for the pilot study was log reaction time. An ANOVA found a main effect for hand (right vs. left),  $F(1, 5) = 84.8$ ,  $p < 0.01$ , but not for interval direction (up vs. down),  $F(1, 5) = 5.9$ ,  $p > 0.05$ , and the interaction term was not significant,  $F(1, 5) = 0.1$ ,  $p > 0.05$ . An analysis of four left-handed subjects found no significant main effects or interaction,  $p > 0.05$ .

Several participants in the pilot study complained that it was difficult to keep the ‘same’ response separate from the ‘up’ and ‘down’ responses. For the main experiment subjects were told to indicate ‘up’ and ‘down’ with the index and middle fingers of one hand, and ‘same’ with the other hand. Subjects were allowed to use whichever hands felt more comfortable. An ANOVA found this change had no effect on log reaction times to unisons. There was a main effect for unison (vs. non-unison),  $F(1,35) = 220.3$ ,  $p < 0.01$ , but not for hand configuration (pilot vs. main experiment),  $F(1,35) = 1.1$ ,  $p > 0.05$ , and the interaction term was not significant,  $F(1,35) = 0.1$ ,  $p > 0.05$ .

Reaction times were recorded by the experiment software. Subjects had only as long to respond as the duration of the note, which led to two forms of biased responses. The majority of notes were 1000ms in duration, but for longer notes subjects had longer

to respond. Longer durations in melodies often correspond to points of metrical stress (Longuet-Higgins & Steedman, 1971), and hence attentional focus (Jones & Yee, 1997), so there was a possibility of uncontrolled bias due to conventions of melodic structure.

It was also possible for subjects to mistakenly enter a response after the next note began. Rather than being entered as a long reaction time, this would register as an extremely short reaction time for the following note. This kind of early response was also possible when subjects guessed what the next note would be rather than waiting to respond until after they had heard it. In order to minimize the effects of late responses and guesses, all responses under 100ms were excluded.

Subjects could respond at any time before the start of the next note. Up to 1000 game points were awarded for fast responses, but points were deducted for responses taking longer than 1 second. A 1000-point penalty was given for no response. To strongly discourage guessing, an incorrect response resulted in a loss of all points earned to that point in the round.

In order to ensure that subjects were attending to the stimuli as melodies rather than simply as local contour changes, each melody was followed by a “bonus round.” The same melody was played again at twice the original tempo (120 eighth notes per minutes, or 500ms IOI). The melody was either exactly the same, or two adjacent pitches were swapped. For 1000 bonus points, subjects were asked to indicate whether the melody was the same or different from the original. The answers to this component of the experiment were not used in the analysis.

Before the experiment, subjects were given a training session. Each subject was supervised through two complete rounds, and asked to continue practicing until comfortable. During the training session explicit visual feedback was provided indicating which answer had been entered and whether it was correct. The only visual feedback given during the actual experiment was scoring information. During both the practice and experiment, most subjects reported a preference to not watch the screen.

### Results

Reaction times were analyzed only for correct responses. Out of a possible 335 data points, this resulted in an average of 309 data points per subject, for a total of 8362 observations. As mentioned earlier, reaction times greater than 1000ms or less than 100ms were excluded.

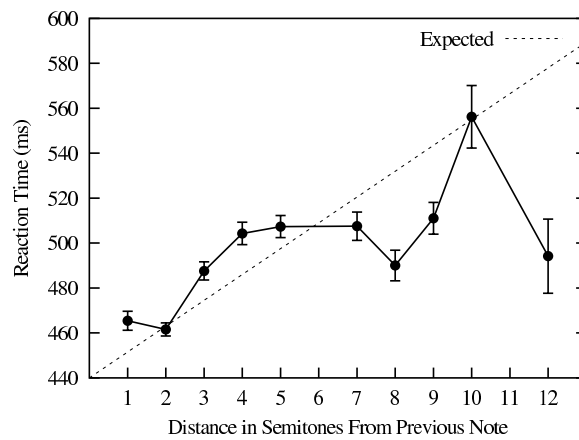
The planned analysis was a multivariate regression, but it is instructive to first examine the individual main effects separately. Because the use of melodic phrases as stimuli creates an unbalanced design, it is impossible to do multivariate ANOVAs. Rather than conducting a separate one-way ANOVA for each variable, qualitative observations will be made instead.

### Univariate Main Effects

The two predictors taken from Schellenberg's reduced model were pitch-proximity and pitch-reversal. The pitch-proximity predictor associates increased pitch distance with

decreased expectancy. The actual observed reaction times for the categories of pitch proximity are shown in Figure 4.2. A trend line is included to show a rough approximation of the expected slope for the variable.

Generally speaking, the overall trend follows the expected path. The reaction times unexpectedly plateau between 3 and 9 semitones (a minor third and a major sixth), and responses after octaves are as fast as any but those after the step intervals. Octaves represented less than 1% of all intervals in the stimuli, however, and were not likely to affect the fit of the model.



**Figure 4.2.** The observed reaction times at the various levels of the pitch-proximity variable. The expected trend of the variable is shown as a dotted line.

The pitch-reversal variable is much more complicated, but the mapping between its values and its predictions are straightforward: higher values should indicate higher expectancy. The actual observed reaction times are shown in Figure 4.3, again with an approximate trend line.

The observed reaction times depart somewhat from the expected values. The three middle categories appear to engender roughly equal levels of expectancy, and the supposed highest category of expectancy generation (2.5) is roughly as slow as the lowest. It is important to note that these values were not derived from Narmour’s theory, but rather through Schellenberg’s own reduction of four parameters to two, which were then weighted by a principle-components analysis and summed.

The “2.5” level of the variable corresponds to large intervals followed by large intervals in the opposite direction. Based on the pitch-proximity predictor, we would expect combinations of large intervals to be relatively unexpected. It may be the particular values Schellenberg assigned to pitch-reversal are only useful when pitch-proximity is taken into account.

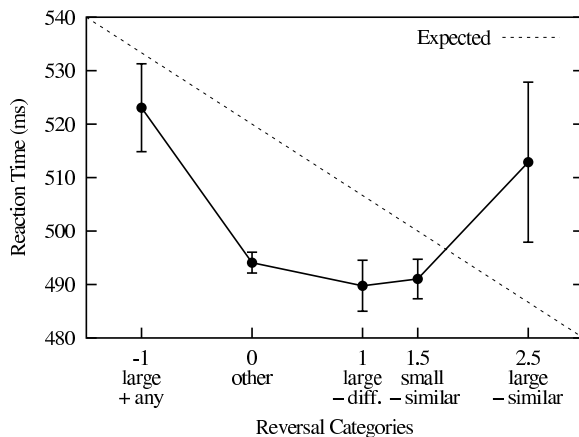
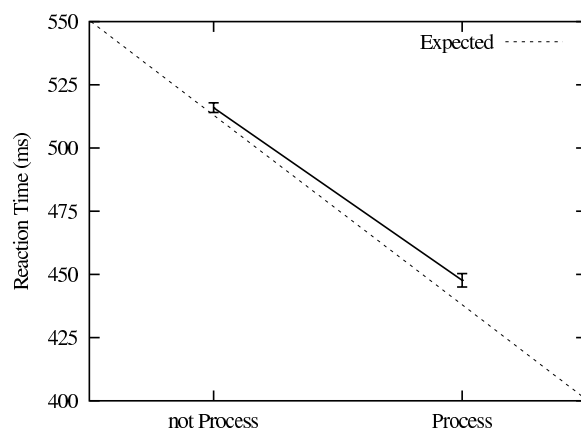


Figure 4.3. The observed reaction times at the five levels of the pitch-reversal variable. The expected trend of the variable is shown as a dotted line. The category labels indicate the size of the first and second intervals; a ‘+’ indicates continuation, and ‘-’ indicates reversal.

The process variable is binomial, and predicts simply that reaction times will be faster after process-like configurations than in other situations. As noted in Chapter 2, the definition of “process” was taken directly from Narmour. The observed reaction times are shown in Figure 4.4, accompanied by an approximation of the expected trend line. The data appear to uphold the predictions fairly well.

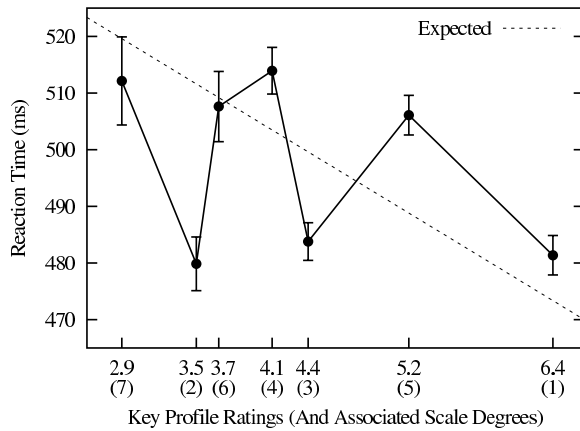


**Figure 4.4.** The observed reaction times at the two levels of the process variable. The expected trend of the variable is shown as a dotted line.

The final variable of interest in the model is the tonality predictor, the major key profile. The observed reaction times for the seven diatonic categories of this variable are shown in Figure 4.5, along with an approximate expected trend line. The categories are shown arranged according to the values of the predictor, with the associated scale degree underneath.

Unfortunately, the correspondence between reaction times and the expected values of the tonality variable is less than obvious. Reaction times were fast to the tonic (scale

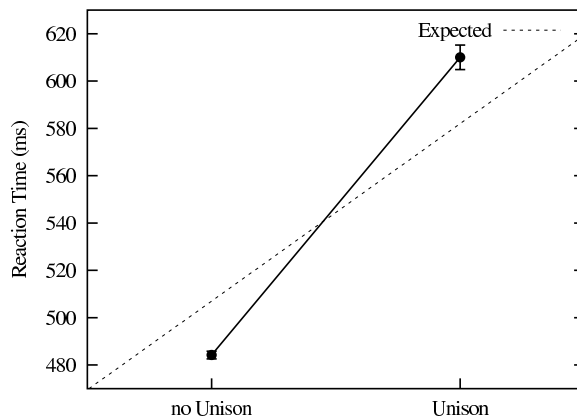
degree 1) and slow to the leading tone (scale degree 7), but there is no even gradation between them. As noted earlier, proximity might be having a stronger effect than tonality, so it may not be useful to consider tonality alone. Nevertheless, there are notable differences in reaction time among the levels of the variable.



**Figure 4.5.** The observed reaction times at the seven diatonic levels of the key-profile variable, ordered according to value (rather than scale degree). The expected trend of the variable is shown as a dotted line.

Finally, the unison covariate predicts that reaction times should be slower in response to unisons (repeated pitches). Indeed, as shown in Figure 4.6, reaction times were considerably slower after unisons.





**Figure 4.6.** The observed reaction times at the two levels of the unison variable. The expected trend of the variable is shown as a dotted line.

Although not all the expected trends were observed when the variables were considered singly — especially with regard to the key profile — the results appear close enough to warrant continuing with the analysis.

### Multivariate Analysis

The planned analysis was a multivariate regression using reaction time as the dependent measure. The independent variables in the model were the three I-R predictors (pitch-proximity, pitch-reversal, and process), key profile, and the unison covariate. To account for the unbalanced within-subjects design, a multilevel regression analysis was conducted using a random intercept (Singer, 1998). Specifically, the intercept was estimated as a latent variable with between-subject variance. A comparison to a model

with no random intercept parameter (the equivalent of a normal OLS — ordinary least squares — regression) found that the deviance ( $-2 \log$ -likelihood) of the multilevel model was significantly better,  $\chi^2(1) = 2178.8$ ,  $p < 0.01$ .

The assumptions of the random intercept model are that both observational and group-level residuals are randomly distributed, and the group variances are homogenous (Snijders & Bosker, 1999). A graphical examination of the estimated subject-level intercept residuals verified that they were normally distributed. Some notable departures from normality were observed in the observation-level residuals, however. Applying a log transform appeared to normalize the residuals of slower responses, but fast responses were generally faster than the model predicted.

Faster responses could possibly be the result of either late responses to a previous note or outright guesses. Either of these would result in a higher proportion of incorrect responses. A plot was constructed to examine the relationship between accuracy and reaction time (Figure 4.7). Reaction times were binned over intervals of 50 milliseconds. (These are labeled according to the lower bound.) For reaction times over 200 ms, accuracy ranged from 90% to 100%, but under 200 ms the accuracy ranged from 50% to at most 80%.

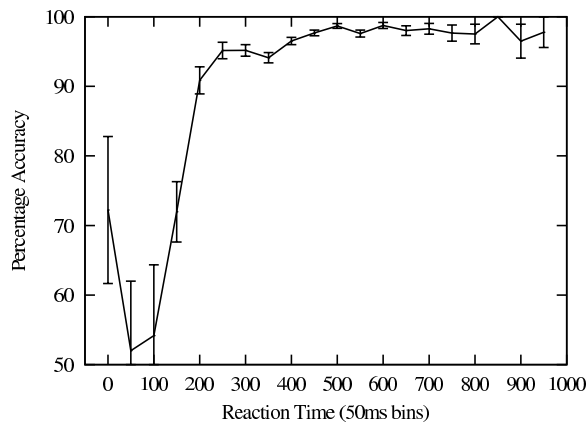


Figure 4.7. Percentage accuracy plotted against reaction time. Reaction times were binned over intervals of 50 milliseconds.

When all 89 observations with reaction times under 200ms (1% of the sample) were deleted, this resulted in residuals that appeared normally distributed.

A Levene test for homogeneity of variance found that there were significant differences in variance across subjects,  $F(26, 8246) = 7.44, p < 0.01$ . One way to compensate for this violation is to add separate variance parameters for each subject (Milliken & Johnson, 1984). The number of covariance parameters in the model jumped from 2 to 28 as a result, but even when taking that into account the deviance of the new model was significantly smaller,  $\chi^2(26) = 292.7, p < 0.01$ .<sup>11</sup>

<sup>11</sup> All multilevel regression analyses were conducted using PROC MIXED in SAS version 8. The final model specification looked like:

```
proc mixed data=expl ic covtest;
  class subject;
  model logRT = proximity reversal process unison / solution;
  random intercept / subject=subject;
  repeated / group=subject;
run;
```

The standardized weight estimates, standard error of the estimates, and *p*-values for all predictors were determined using the REML (residual maximum likelihood) method. The results are shown in Table 4.1. Although the dependent variable for all analyses was log reaction time, effect sizes have been translated back to millisecond reaction times for ease of interpretation.<sup>12</sup>

Predictor	$\beta$	SE	Effect size
I-R model			
pitch-proximity	0.048**	0.011	32ms
pitch-reversal	-0.062**	0.010	44ms
process	-0.204**	0.011	59ms
Tonality model			
key profile	-0.013	0.009	6ms
Covariate			
unison	0.198**	0.010	117ms

\*\* *p* < 0.01.

Table 4.1. Regression Analysis of the I-R and Tonality Models on Log Reaction Time

In OLS regression, model fit is usually summarized as the proportion of variance explained by the model. In the current multilevel model, residual variance is partitioned into subject-level intercept variance and observation-level residual variance. The goal of the model is not to explain the subject-level variance (i.e., how one subject's overall

<sup>12</sup> The estimates of the effect sizes were calculated by multiplying the (non-standardized) weight estimate for each variable by its maximum and minimum values, adding those to the intercept estimate (6.16, or 494 ms), converting to milliseconds, and subtracting one from the other.

mean reaction time varies from another subject's). Rather, it more useful to compare the amount of residual variance in an empty model (which still has random effects) to one that includes all the fixed-effect predictors (Singer, 1998, p. 332). Technically, the residual variance was further partitioned into separate estimates for each subject in order to account for heteroscedasticity, but in order to calculate explained variance the separate estimates are dropped from the model.

An analysis using the full model provides estimates for both the residual variance and the intercept variance. If the intercept variance is fixed at the full model's estimate, then all the predictors are dropped from the model and the analysis is run again, the new residual variance estimate represents the total amount of variance that can be explained by the predictors. In this case, the difference between the empty model estimate (0.0652, the total amount explainable) and the full model estimate (0.0572) is 0.008, which is 12.2% of the total variance. This would be a low percentage of explained variance for most psychological studies, but there is a considerable amount of measurement error in reaction times, especially for decisions requiring as much mental effort as the present task does. Regardless, it will be more useful to compare deviance values than explained variance when evaluating models.

### Discussion

All three of the I-R model predictors were significant and had the expected signs, as did the unison covariate. Smaller pitch proximity led to faster reaction time ( $\beta = 0.048$ ), lower pitch reversal ratings corresponded to higher reaction times ( $\beta = -0.062$ ), responses

were faster in response to process ( $\beta = -0.204$ ), and unisons required more time to identify ( $\beta = 0.198$ ). It is perhaps surprising that the largest standardized weight estimate was assigned to process. As noted earlier, Schellenberg had eliminated a preference for process in his simplified models. These findings suggest that a complete I-R model should include process.<sup>13</sup>

The biggest surprise of the analysis is that the estimate for key profile, the tonality predictor, was insignificant,  $p > 0.05$ . This finding is contrary to previous studies (Cuddy & Lunney, 1995; Schellenberg, 1996; Thompson et al., 1997). It is possible that the reaction time design is insensitive to tonal expectancies in music, or that the particular values of the tonal predictor do not correspond to reaction times in this experiment.

If there were differences in reaction times across scale degrees, then the key profile predictor should be significant when treated as a categorical variable. When the analysis was run again using a set of dummy variables to represent levels of the key profile, the results indeed showed significant differences across categories of the key profile predictor,  $F(6, 7427) = 4.76, p < 0.01$ .<sup>14</sup>

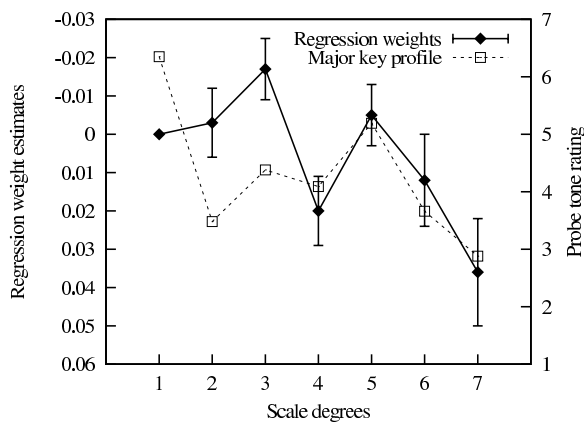
The unstandardized weight estimates for the levels of the categorical key profile predictor are shown in Figure 4.8, measured as deviations in log reaction time. Because

---

<sup>13</sup> One alternate explanation is not accounted for here. Research in visual tasks has shown that the “repetition effect” is especially strong for speeded reaction-time tasks (Bertelson, 1961; Hyman, 1953). That is, a response is facilitated if it is a repetition of the previous response. The speeded design may be magnifying the decision repetition effect; however, that does not preclude the presence of an expectancy repetition effect.

<sup>14</sup> Because the assumption of homoscedasticity was rejected in this analysis, the degrees of freedom were calculated using Satterthwaite’s formula (Littell, Milliken, Stroup, & Wolfinger, 1996).

all levels cannot be estimated simultaneously, the tonic was fixed at 0 as a reference point. As a contrast, the major key profile is also included in the graph, measured on a 7-point Likert scale on the right ordinate axis. To facilitate comparison, the left ordinate is reversed. (A low reaction time and a high Likert scale rating both indicate facilitated expectancy.)



**Figure 4.8.** Scale degree weight estimates (displayed on a reverse ordinate), and the key profile ratings ( $r = -0.53$ ).

Although there are similarities between the two distributions, the correlation is non-significant,  $r = -0.53$ ,  $p = 0.22$ ,  $N = 7$ . This is reflected in the absence of some of the characteristic features of the key profile. Most notably, the tonic was not the most expected scale degree. Comparisons of the least-squares means of the tonic chord members showed that the estimate for scale degree 3 was significantly lower (faster) than

scale degree 1,  $t(7406) = 2.19$ ,  $p < 0.05$ , but scale degrees 5 and 1 were not significantly different,  $t(7390) = 0.57$ ,  $p = 0.57$ . For the major key profile, in comparison, the rating of scale degree 1 is significantly higher than both 3 and 5 (Krumhansl & Kessler, 1982).

If the values of the key profile are not a good match for the reaction time residuals, an important question is, what do the values of the key profile represent? The best explanation of the psychological origins of the probe-tone key profiles is that they are learned zero-order probabilities for scale degrees (Krumhansl, 1990). Evidence in support of the view that key profiles are learned has been found in non-Western cultures (Castellano, Bharucha, & Krumhansl, 1984).

Some of the zero-order probability distributions typically cited in the literature are those of Youngblood (1958) and Knopoff and Hutchinson (1983), shown in Figure 4.9. Because only diatonic scale degrees are used in the present study, only diatonic scale degrees are shown in the figure.



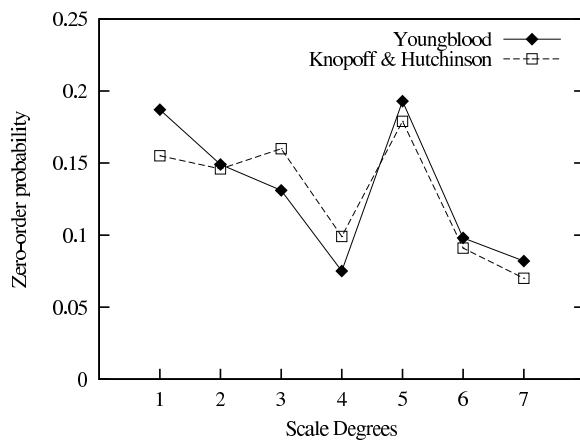


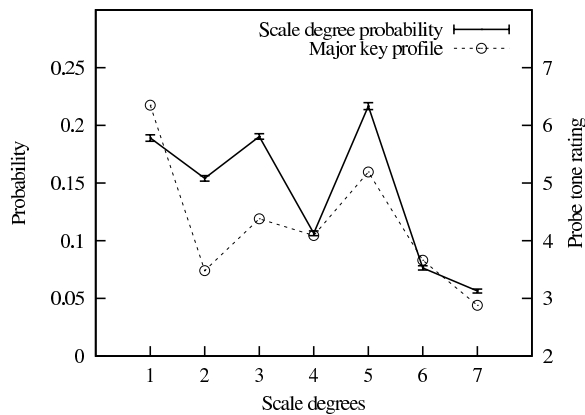
Figure 4.9. Zero-order probabilities of diatonic scale degree occurrence in dozens of Western art song melodies, as measured by Youngblood (1958), N = 2668, and Knopoff and Hutchinson (1983), N = 25,122.

In both the Youngblood (1958) and Knopoff and Hutchinson (1983) studies, all melodic notes were transposed to a common key and averaged into a single distribution. To simplify the problem of finding a common key, the original key for each melody was determined from the key signature set by the composer. In the repertoire used in both of these studies, namely nineteenth-century art song, melodies rarely remain in the same key for the entire song, and new keys are sometimes explicitly notated by a change of key signatures. Youngblood chose to assume that the original key signature was correct throughout, whereas Knopoff and Hutchinson observed all notated changes of key. Although the latter study was probably more accurate as a result, both of these techniques leave open the possibility that notes from different keys were being conflated.

In an attempt to develop a more accurate estimate of scale degree probabilities, a set of 1000 songs was sampled from the Essen Folksong Collection. Folksongs are relatively

short compared to most art songs, which limits the likelihood of modulating to remote keys. It is also more typical of the western European folksong style to remain in the original key (Nettl, 1973).

Notes were converted to scale degrees with the Humdrum Toolkit (Huron, 1993) in order to derive their probabilities. A separate distribution was determined for each melody, and the mean probabilities for the melodies was calculated. The resulting tone profile, shown in Figure 4.10, has a striking similarity to the Knopoff and Hutchinson results in Figure 4.9. (As noted earlier, the Knopoff and Hutchinson sample is less likely to suffer from errors of key misattribution than the Youngblood sample.) The correlation between the two distributions is near-perfect ( $r = 0.99$ ). The agreement between these two independent samples suggests that the Essen distribution is a reasonable estimate of the zero-order probability of scale degrees in traditional Western melodies. For comparison, the major key profile is also plotted in Figure 4.10, with Likert scale ratings on the right ordinate axis.



**Figure 4.10.** The estimated zero-order probabilities of diatonic scale degrees for a major-key folksong from the Essen Folksong Collection. Values were averaged from 1000 folksongs,  $N = 49,265$ . The major key profile is shown for comparison, using a 7-point Likert scale on the right ordinate axis.

The probabilities of scale degree occurrence are similar to the key profile, but the correspondences are not perfect. The highest levels of the proposed tonal hierarchy are less apparent in the probability distribution. The tonic triad members (scale degrees 1, 3, and 5) appear to have high values (although the second scale degree is remarkably prominent), but the tonic is not the most common scale degree.

The original hypothesis justifying the tonality predictor was that reaction times can be predicted by key profile ratings. Referring back to the theory forwarded by Krumhansl (1990) that tonality is learned from the distribution of pitches in music, a more direct hypothesis might be that reaction times can be predicted by the actual zero-order probabilities of scale degrees.

In order to test this alternative hypothesis, another multilevel regression analysis was conducted to consider whether scale degree probability is a better predictor of

reaction times. The zero-order probability of scale degrees was added to the previous model as another tonality predictor. The values for this predictor are those shown in Figure 4.10. The results of this analysis are shown in Table 4.2. The proportion of explained variance for the model was 12.3%.

Predictor	$\beta$	S.E.	Effect size
I-R model			
pitch proximity	0.052**	0.01	38ms
pitch reversal	-0.059**	0.01	44ms
process	-0.206**	0.01	61ms
Tonality model			
key profile	0.003	0.01	2ms
probability	-0.044**	0.01	23ms
Covariate			
unison	0.196**	0.01	119ms

\*\*  $p < 0.01$ .

Table 4.2. Results of the I-R/Tonality Model Including the Probability Predictor.

Again, the weight estimate for key profile was not significant. The estimate for the scale degree probability coefficient was significant, however. The weights of the other predictors did not change much from the first analysis (Table 4.1), which suggests the effects of scale degree probability are relatively independent.

The relative importance of the two tonality predictors was formally tested by comparing nested models. Specifically, the deviance of the full model was compared to

that of the model omitting either the key profile or the probability variable. Estimates in this case were calculated using the standard maximum likelihood algorithm, because the REML method cannot be used to compare fixed effects (Snijders & Bosker, 1999). Dropping the key profile from the full model resulted in no significant change in deviance,  $\chi^2(1) = 0.11$ ,  $p = 0.74$ . The change was significant when scale-degree probability was dropped from the full model, however,  $\chi^2(1) = 21.66$ ,  $p < 0.01$ .

The scale degree probabilities and estimated scale degree weights are shown together in Figure 4.11. Weight estimates are measured in log reaction time on the left ordinate axis, and probability is shown on the right ordinate. As implied by the regression model, the probability profile is more similar to the residuals than was the major key profile. The correlation between the two profiles is significantly above chance,  $r = -0.87$ ,  $p < 0.05$ ,  $N = 7$ .

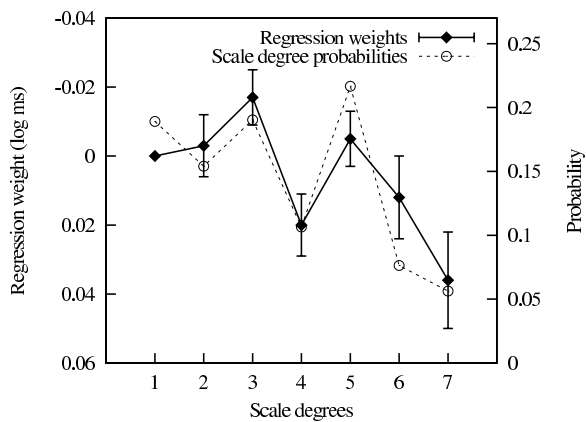


Figure 4.11. Scale degree weight estimates (displayed on a reverse ordinate), plotted with the zero-order probability of scale degrees ( $r = -0.87$ ).

## Overview

In the initial analysis of the data from Experiment 1, it was found that all four I-R model predictors were significant and had the expected signs. Of the four, process had the largest standardized weight. This latter result agrees with the findings of von Hippel (2002), and indicates that the reduced models proposed by Schellenberg (1996; 1997) might be improved by incorporating process expectancies.

The original tonality predictor, the probe-tone key profile, was not found to contribute significantly to the model. In contrast, the actual scale degree probabilities were a significant predictor when added to the model. The question is raised, therefore: why is it that the key profiles perform poorly in this model even though they have predicted tonality quite well in previous studies? Also, if key profiles are learned by exposure to the difference scale degrees, what accounts for the discrepancies between the key profile and the frequency of scale degrees?

One of the motivations for conducting the present study was to remove the potential confound of retrospective perception. It was proposed earlier that the probe-tone method (from which the key profiles were derived) encourages listeners to hear the tone being tested as occurring in a phrase-final position. This may be important, since musical phrases typically end with harmonic “cadences” or with stereotyped melodic figures. Expectancies for phrase-final notes might differ from those of other notes in melodies. Is it possible that the distribution of scale degrees in phrase-final positions is more similar to the key profile?

To answer this question, the distribution of phrase-final scale degrees was calculated for each of the same 1000 Essen folksongs used earlier. The distributions were averaged to estimate the phrase-final scale degree probabilities. Both the resulting distribution and the major key profile are shown in Figure 4.12.

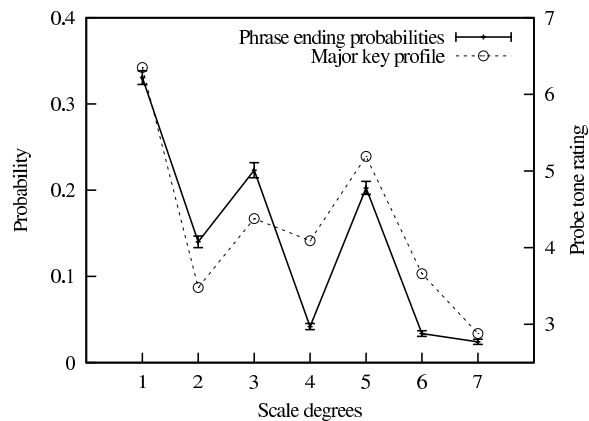


Figure 4.12. The estimated zero-order probabilities of phrase-final scale degrees for major-key folksongs from the Essen Folksong Collection. Values were averaged from the notes in 1000 folksongs,  $N = 5832$ . The major key profile is included as a comparison ( $r = 0.87$ ).

Whereas the correlation between the key profile and the all-note probability distribution was not significant, the correlation between the phrase-final scale degree zero-order probabilities and the major key profile is significant,  $r = 0.87$ ,  $p < 0.05$ ,  $N = 7$ . The characteristic features of the key profile can also be seen in the phrase-final distribution: the tonic is the most common note, followed by degrees 3 and 5, and then the remainder of the scale degrees.

These findings are consistent with the idea that listeners have multiple kinds of tonal expectancy — one for notes that are heard in phrase-final positions, and another for the remainder of the melody. The lack of correspondence between key profiles and scale degree distributions has been a long-standing problem (Butler, 1989a; Parncutt, 1989), so the close correspondence of the phrase-final distribution is striking.

The evidence presented in support of this hypothesis is circumstantial, however. Is it possible, for instance, that the current method and the probe-tone method are measuring fundamentally different things? Or is it that differences in reaction times among scale degrees are heavily laced with measurement error, notwithstanding their significant correlation with zero-order probabilities?

If the claim were to be made that the two methods are measuring different types of expectancy, it would be useful to demonstrate that the probe-tone results can be replicated with the reaction-time method, under the proper circumstances. Experiment 2 was designed explicitly to attempt this demonstration.



## CHAPTER 5

### EXPECTANCY FOR CLOSURE (EXPERIMENT 2)

In the previous experiment it was shown that actual scale-degree probabilities could predict reaction times where the major key profile did not. It was suggested that the probe-tone method (which key profiles are derived from) measures listeners' expectancies only for phrase-final notes. The present experiment is an attempt to measure reaction times only to notes listeners perceive as being phrase-final.

The two features of the probe-tone that have been singled out as problematic are that, first, listeners are always responding to the last note of a sequence; and second, at the time they respond they know the sequence has ended. The hypothesis being forwarded here is that the key profile should be very effective at predicting reaction times under these conditions.

#### Methods

In probe-tone experiments listeners can interpret the probe tone as the last note of a phrase before responding because the music stops after the probe tone. For the reaction time method it is necessary for listeners to know the note will be the last before it begins.

To accomplish this, a counter was added to the interface of the experiment indicating the number of notes remaining in the melody. Subjects were instructed to respond only to the final note at the end of the countdown.

Both of the salient features of the probe tone method were replicated by asking subjects to respond after the counter reached zero. First, subjects were responding only to the final note, and second, they were aware it was the final note before responding. If either or both of these features are critical to the pattern of responses in probe tone experiments, then reaction times to scale degrees in the current experiment should more closely resemble the major key profile.

### Stimuli

A new set of melodic phrases were sampled from the Essen Folksong Collection in a manner similar to that of Experiment 1. The same criteria were used to select a subset of the melodic phrases, except that phrases were not required to contain a large jump. This requirement was dropped both because a larger subset was desirable, and large leaps are rare at phrase endings: less than 5% of all melodic phrases in the Essen Folksong Collection end with leaps of 7 semitones or more. Melodic phrases in which the penultimate and last notes were the same (unisons) were excluded.

Eighty-two melodies were sampled from the resulting subset. The sample was constructed so that the distribution of phrase-final scale degrees was the same as the distribution of all scale degrees in Experiment 1. Sampled phrases are shown in Figure 5.1 for each of the possible scale degree endings.








SD ending	Melody
1	
2	
3	
4	
5	
6	
7	

Figure 5.1. Seven of the 82 melodic phrases used in Experiment 2. One melody is shown for each diatonic scale degree (SD) ending, 1 through 7.

As in Experiment 1, the original keys and ranges were retained for each phrase. The melodies were played at a tempo of 120 eighth notes per minute (500ms per eighth note), double the tempo of Experiment 1. Each melody was preceded by a four-chord progression to establish the key of the melody, again at 120 chords per minute. A 1000ms silence was inserted between the progression and the melody.

### Subjects, Equipment, and Procedure

The subjects were 16 second-year undergraduate music students who participated in the experiment for academic course credit.

As in Experiment 1, the experiment was presented to subjects as a game of skill. Subjects were shown an on-screen counter that indicated how many notes were remaining in the melody. They were instructed that as soon the counter reached zero — as soon as the last note started — they should determine whether the last note had moved up or down from the previous note. (Melodic phrases ending in unisons were not used as stimuli.) Scoring was identical to that in Experiment 1, and subjects were encouraged to respond as quickly as possible.

Before the experiment, subjects were given a training session. Each subject was supervised through two complete rounds, and asked to continue practicing until comfortable. The only visual feedback given during both the training session and the actual experiment was scoring information. The equipment used was identical to that in Experiment 1.

## Results

Only correct responses were analyzed. As in Experiment 1, reaction times under 100ms or over 1000ms were excluded. Out of a possible 82 data points, this resulted in an average of 74 data points per subject, for a total of 1177 observations.

### Univariate Main Effects

The first of the two reduced I-R model variables was pitch-proximity. As before, the mean reaction times were plotted for each level of the predictor (Figure 5.2). Unlike the plot in Experiment 1, there is very little resemblance of the data to the expected trend.

There appear to be differences across levels of the variable, but the range of differences is smaller than in Experiment 1. Perhaps pitch-proximity is a smaller factor in these conditions.

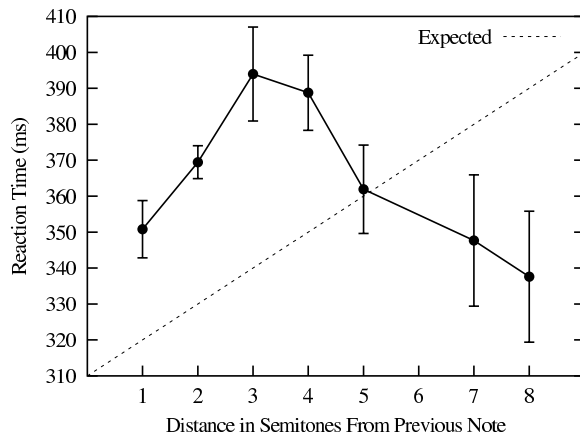


Figure 5.2. The observed reaction times (solid line) at the various levels of the pitch-proximity variable. The expected trend of the variable is shown as a dotted line.

Three levels of the pitch-reversal variable related to large intervals simply do not occur at the ends of the phrases, namely  $-1$ ,  $1$ , and  $2.5$ . The two levels that are found,  $0$  (other) and  $1.5$  (intervals followed by reversals of similar size), are plotted in Figure 5.3. Unfortunately, the results appear to run counter to prediction. At phrase endings, without accounting for other variables, reversals are evidently less expected than other combinations of intervals.

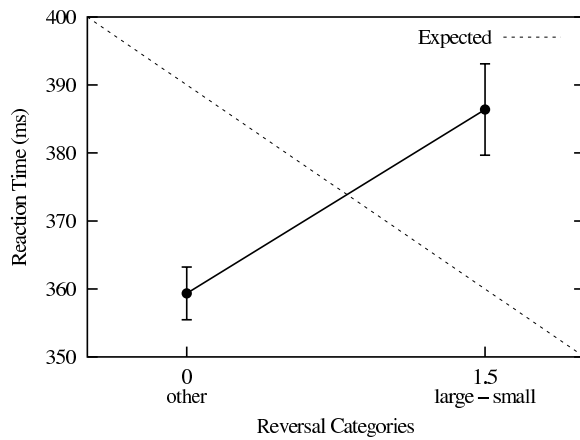


Figure 5.3. The observed reaction times at the two levels of the pitch-reversal variable. The expected trend of the variable is shown as a dotted line.

The final I-R variable, process, is plotted in Figure 5.4. As in the previous experiment, listeners responded faster to instances of process than to other patterns. The effect size appears smaller than in Experiment 1, however.

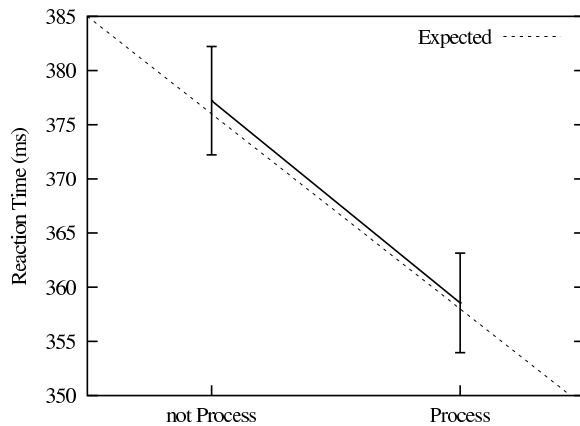


Figure 5.4. The observed reaction times at the two levels of the process variable. The expected trend of the variable is shown as a dotted line.

When the key profile predictor is plotted the results are much closer to the expected trend than in Experiment 1 (Figure 5.5). The most expected scale degrees (the fastest responses) are 1, 3, and 5, just as in the major key profile.

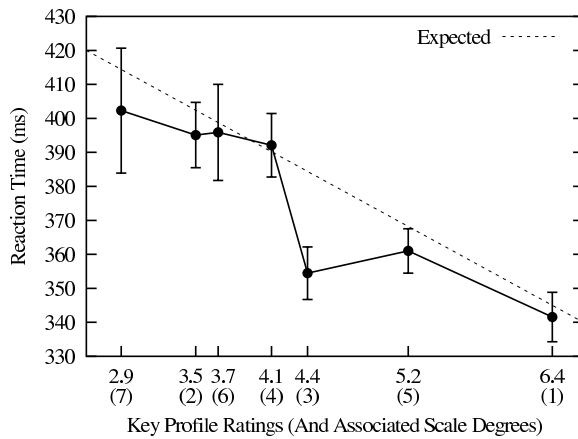


Figure 5.5. The observed reaction times at the various levels of the key-profile variable. The expected trend of the variable is shown as a dotted line.

There are notable departures from the results of the main effects in Experiment 1. Only process showed similar results, and both pitch-proximity and pitch-reversal appear much more muddled. In contrast, the key-profile predictor now appears to follow the expected trend quite closely. Only with the use of a multivariate analysis can we determine if some of these patterns are the results of interactions among predictors.

### Multivariate analysis

A regression analysis was conducted using the three I-R predictors — pitch-proximity, pitch-reversal, and process — to model contour effects (there were no

unisons), and the major key profile to model tonality. A comparison between models found that the addition of a random intercept parameter to the model significantly improved the deviance,  $\chi^2(1) = 427.4$ ,  $p < 0.01$ , so a multilevel model was pursued.

A graphical inspection of the observation-level residuals found no notable deviations from normality. The estimated subject-level intercept residuals also appeared normal. A Levene test for equal variances found significant differences in variance across subjects, however,  $F(15, 1161) = 4.64$ ,  $p < 0.01$ . Adding separate variance parameters for each subject resulted in 15 additional parameters, but even taking that into account the deviance of the larger model was significantly better,  $\chi^2(15) = 103.1$ ,  $p < 0.01$ .

The results of the analysis are shown in Table 5.1. Effect sizes were calculated using the method as in Experiment 1 (see Footnote 12). The proportion of explained variance for the model was 8.9%.

Predictor	$\beta$	SE	Effect size
I-R model			
pitch proximity	0.119**	0.02	95ms
pitch reversal	0.100**	0.03	32ms
process	-0.001	0.03	0ms
Tonality model			
key profile	-0.179**	0.02	64ms

\*\*  $p < 0.01$ .

Table 5.1. Results of the Regression Analysis of Experiment 2

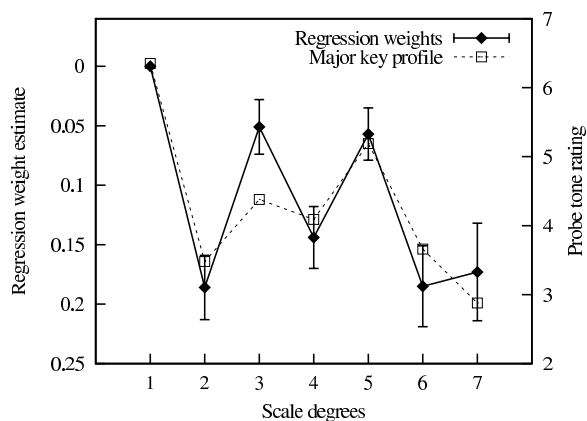


In this analysis, pitch-proximity was significant and had the expected sign. Even though pitch-reversal was significant, however, it had an unexpected sign, and the estimated weight of process was not even significant.

These results turn out to be influenced by a complicated interaction. The only attested levels of pitch-reversal were small-interval reversals (1.5) and “other” interval combinations (0), due to the lack of large intervals. For some unknown reason, reaction times to reversals were slower than to the “other” category, even when the facilitating effects of process were controlled for. In contrast, reaction times after the process archetype were 16ms faster than after other interval combinations. But because pitch-reversal’s “other” category subsumes process, maximizing pitch-reversal obviated the need for process in the model.

The tonality predictor, the Krumhansl-Kessler major key profile, was significant in this analysis and also had the highest standardized weight. This is a departure from the findings of Experiment 1, in which the key profile was not a significant predictor of reaction times.

Again, it is easier to visualize the relationship between the key profile and reaction times by treating the key profile predictor as a categorical variable and plotting the estimated weights of the levels of the variable. As before, the weight of the tonic was fixed at 0 in order to estimate the other levels, as shown in Figure 5.6.



**Figure 5.6.** Scale degree weight estimates in log reaction time (shown on a reverse ordinate axis), along with the major key profile ( $r = -0.91$ ).

Here the characteristic features of the key profile can be seen in the reaction times: the fastest responses were to the tonic, followed by the remaining other tonic triad degrees (3 and 5), and finally the non-tonic degrees. Planned comparisons of the least-squares means confirm this informal observation: reaction times to the tonic were significantly faster than those for both scale degrees 3,  $t(1025) = 2.22$ ,  $p < 0.05$ , and 5,  $t(1039) = 2.57$ ,  $p < 0.05$ . Taken as a whole, the similarities between the two distributions are reflected in a significant correlation,  $r = 0.91$ ,  $p < 0.01$ ,  $N = 7$ .

One of the motivations for this experiment was to explore the striking similarity between key profiles and the scale-degree probabilities for melodic notes in phrase-final position. To verify that both could successfully predict reaction times to scale degrees, phrase-final scale-degree probabilities were added to the model in place of key profile. Like the key profile predictor, the estimate of the phrase-final probability predictor was significant,  $F(1, 1043) = 66.6$ ,  $p < 0.01$ .

## Discussion

The purpose of Experiment 2 was to address the question of whether the new reaction time method is capable of reproducing the results of probe-tone experiments. The most famous of these results are the “key profiles” measured by Krumhansl and Kessler. The hypothesis motivating Experiment 2 was that probe-tone experiments are susceptible to task demands, namely that subjects instinctively hear the last note (“tone”) of the stimulus as a point of closure.

Indeed, the results of Experiment 2 showed that when subjects were aware they were responding to the last note of a melodic phrase, their reaction times were modeled well by the key profile. Furthermore, the scale degrees’ weight estimates replicated the “tonal hierarchy” of the key profile: the fastest reaction times were to the tonic, then scale degrees 3 and 5, followed by the remainder of the diatonic scale degrees.

There are some possible alternate hypotheses. For instance, perhaps these results were not influenced by the counter, but are instead a feature of phrase-final notes under any experimental condition. Fitting the model to only the phrase-final notes of Experiment 1 ( $N = 867$ ) revealed, briefly, that the I-R predictors were all significant,  $p < 0.05$ , but the key profile was not,  $p > 0.05$ . Of course, the picture is more complicated than this. Participants in Experiment 2 often reported they expected the melody to end before the counter reached zero. Structural cues within the melody can apparently create expectations of phrase-finality before the actual ending. Results similar to those in Experiment 2 might be found in the Experiment 1 data if it were possible to select only the responses where each listener expected the phrase to end.

It is also possible, although unlikely, that the property of being the *first* note subjects reply to is the cause of the similarity, but that would presume the prior context does not influence responses. In any case, applying the model to the set of first responses for all melodic phrases in Experiment 1 (N = 940) resulted in only one significant predictor, namely unison,  $F(1, 782) = 158.5, p < 0.01$ .

Earlier it was hypothesized that the results of the probe-tone method are influenced by the effects of perceived closure. Of all the differences between the two experiments presented here, specifically the closure cue and the single response per melody, only the perception of closure appears to explain the commonalities between the key profile and the results of Experiment 2.

#### Further evidence from key-finding

Although key profiles have played a prominent role in psychological studies of music, they have also been of practical application as the basis of a key-finding algorithm (Krumhansl, 1990). The algorithm, designed by Krumhansl and Schmuckler, sums the total duration of each pitch class for the music in question, then determines which of the 12 (enharmonic) major (or minor) keys has the key profile with the highest correlation with the distribution of pitch class durations.

Note that this algorithm calculates a slightly different distribution of notes than those considered earlier. Whereas the distribution in Figure 4.10 represents frequency of occurrence, the algorithm instead uses distribution of durations. It may be that listeners are more aware of durations of notes than they are to the frequencies of their onsets, and a

duration representation may be more useful. This is confounded, however, because long notes (like other accents) tend to occur at strong metrical positions, and listeners devote more attention to stronger metrical positions (Jones & Yee, 1997). Even if duration distributions proved more useful than raw frequency counts, this might be an artifact of auditory attentional mechanisms.

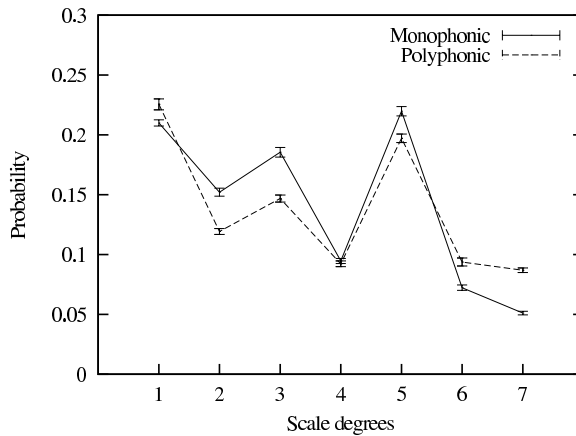
If the key profile is biased toward the probability of scale degrees at phrase endings, perhaps the Krumhansl-Schmuckler key-finding algorithm would benefit by using actual distributions of scale-degree durations derived from music. There are two general forms of Western music to consider here, however. The first is monophonic music, such as the folksong melodies that formed the basis of the experiments reported above. The second is polyphonic music, which may have a different distribution from melodies.

In the process of testing possible improvements to the Krumhansl-Schmuckler algorithm, several different scale-degree distributions were considered as templates. Two psychological measures were used, and two distributions calculated from musical samples. The two psychological measures were the major key profile, and the estimated weights of the scale degrees in Experiment 1. The scale degrees distributions measured were derived from samples of both folksong melodies and polyphonic music.

These four templates were then used to estimate the keys of pieces in order to determine which templates function best under what conditions. Two test samples were constructed, one a sample of folksongs, and the other a sample of polyphonic music.

The duration distributions for folksong and polyphony (shown in Figure 5.7) were derived from a random sample of 1000 major-key folksongs from the Essen Folksong

Collection (the same sample used earlier), and a random sample of non-modulating segments from 250 major-key movements from the MuseData database (CCARH, 2001), respectively. The MuseData database is a collection of thousands of works largely by eighteenth and nineteenth-century European composers. This latter sample was taken from a set in an earlier study that had been constructed to avoid modulations (Aarden, 2001).



**Figure 5.7.** The average distribution of scale degree durations for pieces from two samples. The monophonic sample consisted of 1000 major-key monophonic folksongs from the Essen Folksong Collection (49,265 notes), and the polyphonic sample contained 250 major-key polyphonic segments of movements from the CCARH MuseData database (81,524 notes).

Qualitatively speaking, the duration distributions do not appear noticeably different from the frequency counts in Figure 4.10. The most prominent difference is a slight elevation in the rating of the tonic. Compared to the major key profile shown in Figure 1.1, however, the tonic is still considerably demoted in importance.

Using new samples of 1000 major-key Essen folksongs and 250 major-key segments from the MuseData database, each of the four templates was tested using the Krumhansl-Schmuckler algorithm to determine the percentage of correct key attributions.<sup>15</sup> The results are shown in Table 5.2.

		Correct key identification			
		Experimental measures		Database measures	
Database	N	Key profile	RT weights	Monophonic	Polyphonic
Monophonic	1000	79.0%	87.0%	93.4%	93.7%
Polyphonic	250	80.8%	88.4%	92.4%	94.4%

Table 5.2. Performance of scale degree templates in key identification

The database measures were clearly better than the key profile at correctly identifying keys. A series of chi-square tests found that the key profile template accurately identified keys less often than the templates derived from duration distributions of monophonic music and from polyphonic music,  $p < 0.01$ , but there was no significant difference between the two duration distribution templates.

This finding is trivial in the sense that distributions derived directly from music should obviously be better at describing music than distributions derived indirectly from

<sup>15</sup> This task was easier than the typical problem attempted by key-finding algorithms. Normally the algorithm is forced to consider minor keys as well as major, which would both require the use of a second template and increase the probability of misattribution. In addition, the music being analyzed was selected for its lack of modulations, which are an important confound in key-finding.

psychological experiments. But the reaction time residuals were also better at identifying keys than the key profile for both monophonic music ( $p < 0.01$ ) and polyphonic music ( $p < 0.05$ ), so the measurement error in the key profile is not entirely random.

These tests of the Krumhansl-Schmuckler algorithm reinforce the observation that the results of the new reaction time method correlate better with actual music than the results of the probe-tone method. They also serve to emphasize the point that the relationship between the “key profile” and musical structure is not as straight-forward as is often assumed. The reaction time weights have been presented as better estimates of learned zero-order scale degree probabilities, and this interpretation is consistent with the results in Table 5.2.

### Conclusions

Experiments 1 and 2 together support the hypothesis that the results of the probe-tone method are influenced by the perception of closure. Experiment 2 demonstrates that the reaction time method can reproduce the results of the probe tone method under the special conditions that subjects are expecting to respond to the last note — in other words, at points of closure. Under the normal conditions of the reaction time method, when subjects did not know when the stimuli would end, the results better resembled the scale degree distributions for melodies as a whole.

These two conditions of expectation, the “normal” and “closure-specific,” can be described as perceptual schemata of continuation and closure. The actual probabilities for scale degrees only fit the reaction times for the continuation schema. However, both the



key profile and the probability of phrase-final scale degrees fit the closural schema. Furthermore, the key-finding tests demonstrate that reaction times derived from the continuation schema (Experiment 1) are much closer to distributions in actual music than those generated from closural schema (both the probe tone method and Experiment 2).

Listeners — in this case, undergraduate music students — are evidently capable of learning zero-order scale degree probabilities for both melodies in general and phrase-final notes in particular. There are several motivations for listeners to pay special attention to phrase endings. Phrases are basic units of musical perception (e.g., Jusczyk & Krumhansl, 1993), so phrase endings are important boundary positions. Also, phrase endings are points of low entropy and are extremely formulaic (Manzara, Witten, & James, 1992). These factors contribute to the distinctiveness of phrase endings and provide impetus for the listener to develop a specific schema of perception for phrase endings.

## CHAPTER 6

### RE-EXAMINING THE IMPLICATION-REALIZATION MODEL

The results of the analyses of the I-R model appear to bear out its basic claims, namely that listeners have expectancies for continuations after steps and reversals after leaps. This positive finding is also a negation of one of the I-R variants, however. As noted earlier, the Schellenberg reduced models omit process (step-continuation) as an expectancy. This divergence suggests there may be reason to revisit the question of what elements of the model are redundant.

Attempting to formally test all of the variants would invoke the jeopardy of multiple tests. With that in mind, the results of an informal comparison of the 5-factor parameters model (Krumhansl, 1995; Schellenberg, 1996), the original archetypes model (Narmour, 1990), and the statistical model (von Hippel, 2000) are presented below. No strong conclusions are warranted from any interpretation of the following analyses.

#### The 5-Factor Parameters model

Of all the models of the Implication-Realization theory, the one which has undergone the most experimentation was developed by Krumhansl and Schellenberg in collaboration with Narmour. Narmour's most rigorous definitions of the I-R model were

given in terms of contour archetypes, discussed below. Rather than testing the archetypes directly, however, several factors were derived from theoretical dimensions of expectancy. The resulting 5-factor model (discussed in Chapter 2) was later condensed and subjected to principle-components analysis by Schellenberg to produce his 3- and 2-factor reduced models.

The full 5-factor model was analyzed using the model structure and data set from Experiment 1, but with different fixed effects. As before, the unison predictor was added as a covariate. (Tonality was not used for any tests in this chapter.) The results are shown in Table 6.1. The proportion of explained variance for the model was 11.6%.

Predictor	$\beta$	S.E.	Effect size
I-R model			
registral direction	-0.119**	0.01	34ms
intervallic difference	-0.084**	0.01	26ms
registral return	-0.011	0.01	2ms
pitch proximity	-0.054**	0.01	24ms
pitch closure	0.053**	0.01	24ms
Covariate			
unison	0.203**	0.01	123ms

Note. The data from Experiment 1 were used for the analysis.

\*\*  $p < 0.01$ .







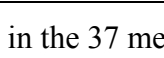
Table 6.1. Regression Analysis of the 5-Factor I-R Model of Krumhansl (1995) and Schellenberg (1996)

The analysis demonstrates that all factors were significant except registral return, although pitch closure had an unexpected sign. The poor showing of registral return is probably due to the redundancy that Schellenberg noted in the model. In the first 3-factor reduced model developed by Schellenberg, both intervallic difference and pitch closure were dropped because of their collinearity with other factors. When those two factors were dropped from the present model, the rest of the predictors remained significant with the expected signs.

Although all archetypes were weighted equally in the registral direction and intervallic difference dimensions as originally defined by Krumhansl and Schellenberg, the results of the experiments in Chapters 4 and 5 have shown that different archetypes may generate unequal levels of expectancy. Experiment 1 demonstrated, for instance, that listeners have a particularly strong expectation for process. This suggests that the factors of the Krumhansl/Schellenberg models can be deceptive, since they collapse multiple archetypes together.

### The Archetype Model

In order to clarify the effects of the individual archetypes relative to one another, a model was created based on Narmour's original description. As described in Chapter 2, the I-R theory has seven prospective archetypes with both implicative and realized components. The survey of these archetypes from Table 2.1 is reproduced below in Table 6.2, including counts of the instances of each from the stimuli in Experiment 1.

Archetype	Implicative	Realized		Diagram	Instances
	Interval	Interval	Direction		
process	Small	Small	Same		106
intervallic process	Small	Small	Opposite		58
registral process	Small	Larger	Same		8
reversal	Large	Smaller	Opposite		38
intervallic reversal	Large	Smaller	Same		11
registral reversal	Large	Larger	Opposite		1
registral return	Any	Similar	Opposite		66

Note. “Instances” indicates the number of each archetype found in the 37 melodies used in the experiment.

Table 6.2. The Prospective Archetypes of Narmour’s I-R Theory, Including Instance Counts

As mentioned in Chapter 2, some variations on the archetypes are theorized to fulfill expectancies better than others. Taking a cue from previous work, however, each archetype predictor was simply dummy coded. The definitions for small and large intervals were adopted from prior studies: intervals smaller than a tritone (6 semitones) were defined as small, and a large interval was defined as larger than a tritone.

One problem with the archetype model is that registral return and intervallic process have a correlation of 0.90 for this sample, which under OLS regression would translate into a Variance Inflation Factor (VIF) of nearly 10. This occurs because most intervals in the sample are small (79%). Most similar-sized reversals will occur for small intervals, which makes the definitions of registral return and intervallic process nearly identical.

This phenomenon was noted by Narmour as a theoretical feature of the model, but it is a statistical hindrance. For reasons of parsimony, registral return was selected to be dropped from the model.

Another problem of note is that there was only a single instance of registral reversal in the sample. A single instance would be a poor test of the predictor, so it was dropped. Registral reversal is defined as a large interval followed by an even larger interval, so its relative scarcity should not be surprising.

The remaining 5-factor prospective archetype model was tested using the data and model structure from Experiment 1. As usual, unison was added as a covariate. Because archetypes represent fulfillments of expectancy, the signs of the regression weights should all be negative (except for unison), indicating a facilitation of mental processing. The results are shown in Table 6.3. The proportion of explained variance for the model was 12.3%.

Predictor	$\beta$	SE	Effect size
I-R model			
process	-0.234**	0.01	68ms
intervallic process	-0.067**	0.01	25ms
registral process	0.021*	0.01	21ms
reversal	-0.071**	0.01	32ms
intervallic reversal	-0.002	0.01	2ms
Covariate			
unison	0.174**	0.01	102ms

Note. The data from Experiment 1 was used for the analysis.

\*  $p < 0.05$ . \*\*  $p < 0.01$ .

Table 6.3. Regression Analysis of the 5-Factor I-R Archetype Model

As the original Experiment 1 analysis had found, process had the largest standardized weight of any of the predictors. The predictors intervallic process, reversal, and unison were significant as well. Although it had a significant weight estimate, registral process had an unexpected sign, and there was no significant effect for intervallic reversal.

On the face of it, this pattern of results appears to have a simple common theme: listeners expect small intervals. The only unexpected archetype was registral process, which ends in a large interval. As a test of this observation, Schellenberg's pitch-proximity predictor was added to the model, but its regression weight was not significant,  $\beta = 0.023$ ,  $SE = 0.01$ ,  $p > 0.05$ , and the effects in Table 6.3 were only mildly attenuated.

Apparently it is not enough to say that listeners expect proximate pitches. Rather, they expect certain types of proximity more than others, particularly those described by process, intervallic process, and reversal.

### The Statistical Model

One alternative to the Implication-Realization models was proposed by von Hippel and Huron (von Hippel, 2000; von Hippel & Huron, 2000). This theory claims that the important features of the I-R models are statistical artifacts of pitch proximity and limited mobility. The clearest example of this is reversal, the tendency for a large interval to be followed by a small interval in the opposite direction. According to the statistical explanation this occurs because, first, large intervals tend to land at the periphery of a melody's pitch distribution, so the most likely continuations will cause a reversal of direction. Secondly, melodies tend to have limited mobility (expressed as a high lag-one autocorrelation), therefore every type of interval is likely to be followed a small interval. Listeners, it is claimed in this theory, are sensitive to the statistical distribution of melodies, and this awareness forms the basis of contour expectancies.

To test the statistical model of pitch proximity, the mean, standard deviation, and lag-one autocorrelation for each melodic phrase were determined, and a value was computed for each note using the formula specified in von Hippel (2000).<sup>16</sup> Again, the model structure and data from Experiment 1 were used, but the only fixed effects were

---

<sup>16</sup> As printed, the formula is missing an autocorrelation term in the denominator (personal communication, von Hippel, 2002).



the statistical predictor and the unison covariate. The results indicated that the single statistical measure was successful at modeling listeners' contour expectancies,  $\beta = -0.058$ ,  $SE = 0.01$ ,  $p < 0.01$ , and had an estimated effect size of 27ms. The unison covariate was also significant,  $\beta = 0.215$ ,  $SE = 0.01$ ,  $p < 0.01$ . The proportion of explained variance for the model was 6.2%.

### Overview

Although there is no statistical means to formally compare these models, it is possible to roughly gauge their relative merits. A common method is to compare measures of model fit that penalize log-likelihood according to the number of fixed and random parameters in the model. The AIC (Akaike Information Criterion) is the most common measure, but the BIC (Bayesian Information Criterion) is more conservative when accounting for the number of parameters (Littell et al., 1996).<sup>17</sup> When calculated from deviance values (as was the case here), lower values are preferred — that is, values that are more negative. The model fit numbers are shown in Table 6.4 for the four models considered so far.

---

<sup>17</sup> The general formula for both the AIC and BIC is

$$deviance + k * npar$$

where *deviance* is  $-2 * \log\text{-likelihood}$ , *npar* is the number of parameters in the model,  $k = 2$  for the AIC, and  $k = \log(n)$  for the BIC, where  $n$  is the number of observations (Sakamoto, Ishiguro, & Kitagawa, 1986).

Model	AIC	BIC	Deviance	Predictors
Experiment 1	-291.9	-249.1	-357.9	4
I-R archetypes	-303.8	-258.5	-373.8	6
Krumhansl/Schmuckler	-249.1	-203.7	-319.1	6
von Hippel	214.0	254.2	152.0	2

Note. Lower (more negative) values of the AIC and BIC are better.

Table 6.4. Model Fit Criteria for the Four Variants of the I-R Model

Of the four, the best bang for the buck comes from the archetype model, since it has both the lowest deviance and the lowest information criteria values. Not all of the predictors were useful in the archetypes model, however: intervallic reversal had an insignificant estimate. This reduces the I-R model to the simple statement that listeners strongly expect small intervals to be followed by small intervals in the same direction — “registral process” is the violation of that expectancy and was the only archetype with the wrong sign — and they also tend to expect small intervals reversing direction.

## CHAPTER 7

### CONCLUSIONS

The experiments presented here tested a new method of measuring melodic expectancy using reaction time. A motivation for the development of this method was to avoid potential confounds of retrospective perception that might affect methods such as the probe-tone measure. A second motivation was to gather prospective expectation measures from controlled stimuli, rather than using methods that rely on subjects' invented continuations for incomplete stimuli.

In Experiment 1 it was found that several features of the Implication-Realization model correlated well with reaction times, especially the archetype known as “process.” It was found that although key profiles are hypothesized to originate in learned scale degree probabilities, there are meaningful differences between the major key profile and the actual scale degree probabilities. The actual probabilities of scale degrees added significantly to the fit of the I-R model, whereas the key profile did not.

Experiment 2 asked subjects to respond only to notes that were known to be the final notes of melodic sequences, in an attempt to replicate the findings of the probe-tone method using reaction times. The results of this study showed that the “key profile” had the highest standardized weight of any of the predictors. In addition, the weight estimates

for each of the scale degrees replicated the hierarchical features of the key profile. As converging evidence, the major key profile was found to be quite similar to the probabilities of scale degrees at phrase endings. This suggests the key profile is learned from the distribution of scale degrees at phrase endings.

A distinction has been drawn between prospective and retrospective measurement methods for expectancy; that is, between measurement methods in which expectancies are related to a forthcoming note (e.g., reaction time and continuation studies), and those in which subjects are asked to report expectancies after the stimulus (e.g., probe tone studies). Prospective expectancies appeared to be more related to actual probabilities of occurrence in music (Experiment 1), whereas expectations of tonal closure (Experiment 2 and the probe tone method) were found to be more similar to phrase-final probabilities.

A great deal of research has been conducted using the probe-tone paradigm, and the findings presented here should not necessarily be seen as questioning the validity of those results. The value of the key profiles has been as a quantification of tonality, and the primary features of tonality are common to both key profiles and actual zero-order scale degree probabilities: that is, all things being equal, diatonic pitches are more expected than non-diatonic pitches. In addition, moments of cadence and closure are the points at which tonality is firmly established (Piston, 1941). To the extent that tonality can be equated with tonal stability, the key profile appears to be a very apt quantifier.

However, the results of this study suggest that the probe-tone method is confounded by the expectation (for many listeners) of closure, and that researchers need to be careful in interpreting the meaning of probe-tone results.

The primary contour expectancy, as operationalized by the Implication-Realization models, was found to be that of process, alternately known as “step momentum.” There were also effects of reversal (sometimes called “post-skip reversal” or “gap-fill”) and intervallic process. Collectively these predictors appear to have a great deal in common with pitch proximity, but when put together in a model the I-R predictors are much more successful at explaining reaction time variance than is pitch proximity. Consistent with Narmour’s claims, some types of proximity are apparently more expected than others.

A number of limitations were imposed on this study for the sake of simplicity. The decision to exclude minor keys was made because the major mode is a simpler construct. It would be a useful extension of this work to demonstrate the same effects for minor-key melodies.

The tempi chosen for the stimuli were selected on the basis of the author’s sense of the fastest comfortable response time. It would be useful to determine what effects if any a change in the time interval between notes might have. Are there peak moments of expectancy, or do different forms of expectation form at different time intervals? These questions await further research.

Some alternate formulations of the Implication-Realization model were considered in Chapter 6, but there are many other components of the Implication-Realization model that still have not been tested. Foremost among these are the retrospective archetypes, which are postulated to shape expectancies once the “prospective” archetypes considered in the present models have been violated. Presumably these expectancies are formed over the course of 4 or 5 notes, rather than the 1- to 3-note spans considered here.

Furthermore, one of the assumptions of the I-R models is that direction is perceived as a purely relative phenomenon. It can be readily observed, however, that in Western music leaps typically ascend, and steps more often descend. If listeners are sensitive to these tendencies, it may be that expectancies for the I-R archetypes interact with absolute direction.

The participants in these studies were all undergraduate music students with years of musical training. An outstanding question is whether sensitivity to scale degree frequencies can be learned from mere exposure to music, or whether active engagement in musical production is necessary. Some studies have demonstrated that key profiles are significantly correlated with the probe-tone responses of non-musicians (Lamont & Cross, 1994), but the number of hierarchical levels of the distribution learned by non-musicians is unclear. A further question of interest is whether a distinct closure schema of listening is something that arises with musical expertise, or if it can occur as a result of unskilled acculturation.

It will be a useful extension to discover what other kinds of expectancies listeners have. Are learned zero-order probabilities really just simplistic reductions of learned first-order or second-order probabilities? Or are they artifacts of a finite number of familiar melodic formulae?

The results of this study demonstrate that the connections between melodic structure and melodic expectancy are more straightforward than has been previously demonstrated. Melodic expectancy is related directly to the distributions of events in music. The results also indicate that melodic expectation is more dynamic than has been

evidenced in much of the literature. Tonal expectancies can vary according to the context or cues that are presented, as when explicit cues of phrase endings are provided. Contour expectancies also vary according to context, again perhaps because of differences in the probability of particular contour archetypes. As more sensitive methods of measurement arise, these findings suggest a more dynamic and detailed picture of melodic expectancy may emerge.

## REFERENCES

- Aarden, B. (2001). *An empirical study of chord-tone doubling in Common Era music*. Unpublished Master's Thesis, The Ohio State University, Columbus, OH.
- Bartlett, J. C., & Dowling, W. J. (1988). Scale structure and similarity of melodies. *Music Perception, 5*(3), 285-314.
- Bertelson, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. *Quarterly Journal of Experimental Psychology, 13*, 90–102.
- Besson, M., & Faieta, F. (1995). An event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception & Performance, 21*(6), 1278–1296.
- Bharucha, J., & Krumhansl, C. L. (1983). The representation of harmonic structure in music: Hierarchies of stability as a function of context. *Cognition, 13*(1), 63-102.
- Bingham, W. V. D. (1910). *Studies in melody*. Baltimore: The Review Publishing Co.
- Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: The MIT Press.
- Brown, H. (1988). The interplay of set content and temporal context in a functional theory of tonality perception. *Music Perception, 5*(3), 219-249.
- Brown, H., & Butler, D. (1981). Diatonic trichords as minimal tonal cue-cells. *In Theory Only, 5*(6-7), 39-55.
- Brown, H., Butler, D., & Jones, M. R. (1994). Musical and temporal influences on key discovery. *Music Perception, 11*(4), 371-407.
- Butler, D. (1982). *The initial identification of tonal centers in music*. Paper presented at the NATO Conference on the acquisition of symbolic skills, University of Keele, England.



- Butler, D. (1989a). Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry. *Music Perception*, 6(3), 219-242.
- Butler, D. (1989b). Response to Carol Krumhansl. *Music Perception*, 7(3), 325–338.
- Butler, D. (1992). On pitch-set properties and perceptual attributes of the minor mode. In M. R. Jones & S. Holleran (Eds.), *Cognitive bases of musical communication*. (pp. 161-169). Washington, D.C.: American Psychological Association.
- Butler, D., & Brown, H. (1984). Tonal structure versus function: Studies of the recognition of harmonic motion. *Music Perception*, 2(1), 6-24.
- Butler, D., & Brown, H. (1994). Describing the mental representation of tonality in music. In R. Aiello & J. A. Sloboda (Eds.), *Musical perceptions*. (pp. 191-212). New York: Oxford University Press.
- Carlsen, J. C. (1981). Some factors which influence melodic expectancy. *Psychomusicology*, 1(1), 12–29.
- Carlsen, J. C., Divenyi, P. L., & Taylor, J. A. (1970). A preliminary study of perceptual expectancy in melodic configurations. *Council for Research in Music Education Bulletin*, 22, 4–12.
- Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of North India. *Journal of Experimental Psychology: General*, 113(3), 394–412.
- CCARH. (2001). The MuseData database (<http://www.musedata.org>). Stanford, CA: Center for Computer Assisted Research in the Humanities.
- Cuddy, L. L., & Badertscher, B. (1987). Recovery of the tonal hierarchy: Some comparisons across age and levels of musical experience. *Perception & Psychophysics*, 41(6), 609-620.
- Cuddy, L. L., & Cohen, A. J. (1976). Recognition of transposed melodic sequences. *Quarterly Journal of Experimental Psychology*, 28(2), 255–270.
- Cuddy, L. L., Cohen, A. J., & Miller, J. (1979). Melody recognition: The experimental application of musical rules. *Canadian Journal of Psychology*, 33(3), 148-157.
- Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception & Psychophysics*, 57(4), 451-462.

- Deutsch, D. (1969). Music recognition. *Psychological Review*, 76, 300–307.
- Dowling, W. J. (1971). Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, 49(2), 524–531.
- Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology*, 1(1), 30-49.
- Eerola, T., Toiviainen, P., & Krumhansl, C. L. (2002). *Real-time prediction of melodies: Continuous predictability judgments and dynamic models*. Paper presented at the 7th International Conference on Music Perception and Cognition, Sydney, Australia.
- Farnsworth, P. R. (1926). *Ending preferences and apparent pitch of a combination of tones*. Unpublished Ph. D. Thesis, The Ohio State University, Columbus, OH.
- Fodor, J. A. (1983). *Modularity of mind: an essay on faculty psychology*. Cambridge, MA: MIT Press.
- Fraisse, P. (1957/1963). *The psychology of time* (J. Leith, Trans.). New York: Harper & Row.
- Francés, R. (1958/1988). *The perception of music* (W. J. Dowling, Trans.). Hillsdale, NJ: Lawrence Erlbaum.
- Helmholtz, H. v. (1877/1948). *On the sensations of tone as a physiological basis for the theory of music* (A. J. Ellis, Trans. 4th ed.). New York: P. Smith.
- Holleran, S., Jones, M. R., & Butler, D. (1995). Perceiving implied harmony: The influence of melodic and harmonic context. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21(3), 737-753.
- Hughes, M. (1977). A quantitative analysis. In M. Yeston (Ed.), *Readings in Schenker analysis and other approaches* (pp. 144-164). New Haven: Yale University Press.
- Huron, D. B. (1993). The Humdrum Toolkit: Software for Music Researchers [Three computer disks and 16-page installation guide]. Stanford, CA: Center for Computer Assisted Research in the Humanities.
- Huron, D. B. (1994). Interval-class content in equally-tempered pitch-class sets: Common scales exhibit optimal tonal consonance. *Music Perception*, 11(3), 289–305.
- Huron, D. B., & Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echo memory. *Psychomusicology*, 12(2), 154-171.

- Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45, 188–196.
- Janata, P., & Reisberg, D. (1988). Response-time measures as a means of exploring tonal hierarchies. *Music Perception*, 6(2), 161-172.
- Jones, M. R., & Yee, W. (1997). Sensitivity to time change: the role of context and skill. *Journal of Experimental Psychology: Human Perception & Performance*, 23(3), 693–709.
- Juszyk, P. W., & Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of Experimental Psychology: Human Perception & Performance*, 19(3), 627–640.
- Knopoff, L., & Hutchinson, W. (1983). Entropy as a measure of style: The influence of sample length. *Journal of Music Theory*, 27(1), 75-97.
- Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, 11(3), 346-374.
- Krumhansl, C. L. (1987). Tonal and harmonic hierarchies. In J. Sundberg (Ed.), *Harmony and tonality*. Stockholm: Royal Swedish Academy of Music.
- Krumhansl, C. L. (1989). Tonal hierarchies and rare intervals in music cognition. *Music Perception*, 7(3), 309–324.
- Krumhansl, C. L. (1990). *The cognitive foundations of musical pitch*. New York: Oxford University Press.
- Krumhansl, C. L. (1995). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, 17, 53–80.
- Krumhansl, C. L., Bharucha, J., & Castellano, M. A. (1982). Key distance effects on perceived harmonic structure in music. *Perception & Psychophysics*, 32(2), 96-108.
- Krumhansl, C. L., Bharucha, J. J., & Kessler, E. J. (1982). Perceived harmonic structure of chords in three related musical keys. *Journal of Experimental Psychology: Human Perception & Performance*, 8(1), 24-36.
- Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4), 334-368.

- Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Jaervinen, T., & Eerola, T. (1999). Melodic expectation in Finnish spiritual folk hymns: Convergence of statistical, behavioral, and computational approaches. *Music Perception, 17*(2), 151-195.
- Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception & Performance, 5*(4), 579-594.
- Lamont, A., & Cross, I. (1994). Children's cognitive representations of musical pitch. *Music Perception, 12*(1), 27-55.
- Lipps, T. (1885/1926). *Psychological Studies* (H. C. Sanborn, Trans. 2nd ed.). Baltimore: Williams & Wilkins.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS© System for Mixed Models*. Cary, NC: SAS Institute Inc.
- Longuet-Higgins, H. C., & Steedman, M. J. (1971). On interpreting Bach. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence* (Vol. 6). Edinburgh: Edinburgh University Press.
- Manzara, L. C., Witten, I. H., & James, M. (1992). On the entropy of music: An experiment with Bach Chorale melodies. *Leonardo Music Journal, 2*(1), 81-88.
- Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology, 26*, 3-67.
- Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.
- Meyer, M. (1901). *Contributions to a psychological theory of music*. Columbia, MO: University of Missouri.
- Milliken, G. A., & Johnson, D. E. (1984). *The analysis of messy data* (Vol. 3). Belmont, CA: Lifetime Learning Publications.
- Mitroudot, L. (2001). Infants' melodic schemas: analysis of song productions by 4- and 5-year-old subjects. *Musicae Scientiae, 5*(1), 83-104.
- Myors, B. (1998). A simple graphical technique for assessing timer accuracy of computer systems. *Behavior Research Methods Instruments & Computers, 30*(3), 454-456.
- Narmour, E. (1977). *Beyond Schenkerism: the need for alternatives in music analysis*. Chicago: University of Chicago Press.

- Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. Chicago: University of Chicago Press.
- Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realization model*. Chicago: University of Chicago Press.
- Nettl, B. (1973). *Folk and traditional music of the western continents*. Englewood Cliffs, NJ: Prentice-Hall.
- Palmer, C., & Holleran, S. (1994). Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *Perception & Psychophysics*, 56(3), 301-312.
- Palmer, C., & Krumhansl, C. L. (1987a). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception & Performance*, 13(1), 116-126.
- Palmer, C., & Krumhansl, C. L. (1987b). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, 41(6), 505-518.
- Parncutt, R. (1989). *Harmony: a psychoacoustical approach*. New York: Springer-Verlag.
- Piston, W. (1941). *Harmony*. New York: W. W. Norton.
- Pomerantz, J. R. (1981). Perceptual organization in information processing. In J. R. Pomerantz & M. Kubovy (Eds.), *Perceptual Organization* (pp. 141-180). Hillsdale, NJ: Erlbaum.
- Povel, D. J. (1996). Exploring the elementary harmonic forces in the tonal system. *Psychological Research-Psychologische Forschung*, 58(4), 274-283.
- Rameau, J. P. (1722/1971). *Treatise on harmony* (P. Gossett, Trans.). New York: Dover.
- Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion statistics*. Hingham, MA: Kluwer Academic.
- Schaffrath, H. (1995). The Essen folksong collection. D. Huron (ed.). [Four computer disks containing 6,225 folksong transcriptions and 34-page research guide]. Stanford, CA: Center for Computer Assisted Research in the Humanities.
- Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realization model. *Cognition*, 58(1), 75-125.

- Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception, 14*(3), 295-318.
- Schenker, H. (1935/1979). *Free composition: volume III of New musical theories and fantasies* (E. Oster, Trans.). New York: Longman.
- Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception, 7*(2), 109-149.
- Shepard, R. N. (1964). Circularity in judgments of relative pitch. *Journal of the Acoustical Society of America, 36*, 2346–2353.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational & Behavioral Statistics, 23*(4), 323–355.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: SAGE Publications.
- Stumpf, C., & Meyer, M. (1898). Maassbestimmungen über die Reinheit consonanter Intervalle. *Zeitschrift für Psychologie, 18*, 321.
- Teplov, B. M. (1966). *Psychologie des aptitudes musicales* (J. Deprun, Trans. from Russian). Paris: Presses Universitaires de France.
- Thompson, W. F. (1996). The analysis and cognition of basic melodic structures: The implication-realization model (review). *Journal of the American Musicological Society, 49*(1), 127-145.
- Thompson, W. F., Cuddy, L. L., & Plaus, C. (1997). Expectancies generated by melodic intervals: Evaluation of principles of melodic implication in a melody production task. *Perception & Psychophysics, 59*(7), 1069–1076.
- Thomson, W. (2001). Deductions concerning inductions of tonality. *Music Perception, 19*(1), 127-138.
- Trainor, L. J., & Trehub, S. E. (1994). Key membership and implied harmony in Western tonal music: Developmental perspectives. *Perception & Psychophysics, 56*(2), 125-132.
- Unyk, A. M., & Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology, 7*(1), 3–23.
- von Hippel, P. (2000). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception, 17*(3), 315-327.

von Hippel, P. (2002). *Melodic-expectation rules as learned heuristics*. Paper presented at the 7th International Conference on Music Perception and Cognition, Sydney, Australia.

von Hippel, P., & Huron, D. B. (2000). Why do skips precede reversals? The effect of tessitura on melodic structure. *Music perception*, 18(1), 59-85.

Youngblood, J. E. (1958). Style as information. *Journal of Music Theory*, 2, 24-35.