Music Perception
Summer 2004, Vol. 21, No. 4, 499–543

Parsing of Melody: Quantification and Testing of the Local Grouping Rules of Lerdahl and Jackendoff's *A Generative Theory of Tonal Music*

BRADLEY W. FRANKLAND

Dalhousie University

ANNABEL J. COHEN

University of Prince Edward Island

In two experiments, the empirical parsing of melodies was compared with predictions derived from four grouping preference rules of *A Generative Theory of Tonal Music* (F. Lerdahl & R. Jackendoff, 1983). In Experiment 1 (n = 123), listeners representing a wide range of musical training heard two familiar nursery-rhyme melodies and one unfamiliar tonal melody, each presented three times. During each repetition, listeners indicated the location of boundaries between units by pressing a key. Experiment 2 (n = 33) repeated Experiment 1 with different stimuli: one familiar and one unfamiliar nursery-rhyme melody, and one unfamiliar, tonal melody from the classical repertoire. In all melodies of both experiments, there was good within-subject consistency of boundary placement across the three repetitions (mean r = .54). Consistencies between Repetitions 2 and 3 were even higher (mean r = .63). Hence, Repetitions 2 and 3 were collapsed. After collapsing, there was high between-subjects similarity in boundary placement for each melody (mean r = .62), implying that all participants parsed the melodies in essentially the same (though not identical) manner. A role for musical training in parsing appeared only for the unfamiliar, classical melody of Experiment 2. The empirical parsing profiles were compared with the quantified predictions of Grouping Preference Rules 2a (the Rest aspect of Slur/Rest), 2b (Attack-point), 3a (Register change), and 3d (Length change). Based on correlational analyses, only Attack-point (mean r = .80) and Rest (mean r = .54) were necessary to explain the parsings of participants. Little role was seen for Register change (mean r = .14) or Length change (mean r = –.09). Solutions based on multiple regression further reduced the role for Register and Length change. Generally, results provided some support for aspects of *A Generative Theory of Tonal Music,* while implying that some alterations might be useful.

Received May 18, 1999, accepted September 10, 2003

Address correspondence to B. W. Frankland, Department of Psychology, Dalhousie University, Halifax, NS, Canada, B3H 4J1. (e-mail: Brad.Frankland@dal.ca)

ISSN: 0730-7829. Send requests for permission to reprint to Rights and Permissions, University of California Press, 2000 Center St., Ste. 303, Berkeley, CA 94704-1223.

LERDAHL and Jackendoff's (1983) *A Generative Theory of Tonal Music* (hereafter *GTTM*) begins at the level of the acoustic sequence of sounds in the musical surface and arrives at a global (cognitive) organization of the piece as music heard by an idealized listener. *GTTM* assumes that "the listener naturally organizes the sound signals into units such as motives, themes, phrases, periods, theme-groups, sections and the piece itself. . . . Our generic term for all these units is *group*" (Lerdahl & Jackendoff, 1983, p. 12). *GTTM* provides a strictly hierarchical parsing of a piece of music using five Grouping Well-Formedness Rules to define the general characteristics of a hierarchy (Lerdahl & Jackendoff, 1983, pp. 37–39). Small groups of notes at the lower levels of the hierarchy are combined to create larger groups at higher levels of the hierarchy. The theory also holds that this grouping hierarchy is independent of (but not isolated from) the metrical hierarchy (Lerdahl & Jackendoff, 1983, p. 12; see also pp. 12–35). As such, it is possible to examine the grouping hierarchy without the complication of metrical analysis.

The focus of the current work is a subset of the seven Grouping Preference Rules (GPRs; Lerdahl & Jackendoff, 1983, pp. 39–67; shown verbatim in Table 1) that control the content and organization of the groups within the grouping hierarchy. The basis for the GPRs, and consequently the grouping structure of the hierarchy, is the phenomenal accent, which is:

> any event at the musical surface that gives emphasis or stress to a moment in the musical flow. Included in this category are attack-points of pitch-events, local stress like sforzandi, sudden changes in dynamics or timbre, long notes, leaps to relatively high or low notes, harmonic changes and so forth. (Lerdahl & Jackendoff, 1983, p. 17)

Ultimately, the grouping structure depends on the detection of change in the pitches, timbre, intensity, and the timings of notes. For the current work, it is important to realize that the roles of "harmonic changes and so forth" were not clarified and that issues of tonality seem to be restricted to the higher level rules. *GTTM* argues that the GPRs predict *possible* boundaries between units of music, thereby informing of *possible* parsings of melody (the terms groups, boundaries, units, and parsing are often used interchangeably). However, of the seven GPRs, only GPRs 2 (Proximity) and 3 (Change) define possible boundaries (see Table 1). GPR 1 simply encourages the avoidance of small groups. GPR 4 (Intensification) seems to address the effect of congruent applications of GPRs 2 and 3. GPRs 5 (Symmetry), 6 (Parallelism), and 7 (Time-Span and Prolongation Reduction) entail global relations between larger segments of the music. As such, GPRs 2 and 3 define possible boundaries while GPRs 1, 4, 5, 6, and 7 determine which of those boundaries are to be retained at higher

TABLE 1
The Grouping Preference Rules (GPRs) as Defined by the Generative Theory of Tonal Music (Lerdahl & Jackendoff, 1983)

Rule 1		Avoid analyses with very small groups—the smaller the less preferable.
Rule 2	Proximity	Consider a sequence of four notes n_1 n_2 n_3 n_4. All else being equal, the transition n_2-n_3 may be heard as a group boundary if:
	a. Slur/Rest[a]	the interval of time from the end of n_2 to the beginning of n_3 is greater than that from the end of n_1 to the beginning of n_2 and that from the end of n_3 to the beginning of n_4.
	b. Attack-point[a]	the interval of time between the attack points of n_2 and n_3 is greater than that between n_1 and n_2 and that between n_3 and n_4.
Rule 3	Change	Consider a sequence of four notes n_1 n_2 n_3 n_4. All else being equal, the transition n_2-n_3 may be heard as a group boundary if:
	a. Register[a]	the transition n_2 to n_3 involves a greater intervallic distance than both n_1 to n_2 and n_3 to n_4.
	b. Dynamics	the transition n_2 to n_3 involves a change in dynamics and n_1 to n_2 and n_3 to n_4 do not.
	c. Articulation	the transition n_2 to n_3 involves a change in articulation and n_1 to n_2 and n_3 to n_4 do not.
	d. Length[a]	n_2 and n_3 are of different lengths, and both pairs n_1, n_2 and n_3, n_4 do not differ in length.
Rule 4	Intensification[a]	Where the effects of Group Preference Rules 2 and 3 are relatively more pronounced, a larger level group boundary may be placed.
Rule 5	Symmetry	Prefer grouping analyses that most closely approach the ideal subdivision of groups into two parts of equal length.
Rule 6	Parallelism	Where two or more segments of the music can be construed as parallel, they preferably form parallel parts of groups.
Rule 7	Time-Span and Prolongation Stability	Prefer a grouping structure that results in more stable time-span and/or prolongation reductions.

NOTE—These definitions are taken directly from Lerdahl and Jackendoff (1983, pp. 45–52).
[a]Rules that were quantified and tested in this work.

levels of the hierarchy (cf. Lerdahl & Jackendoff, 1983, pp. 48–49). GPRs 2 and 3 depend on the phenomenal accent, and GPRs 1, 4, 5, 6, and 7 depend on GPRs 2 and 3. Hence, GPRs 2 and 3 are fundamental to any subsequent grouping structure. This also implies that GPRs 2 and 3 can be analyzed with minimal regard to the rest of the theory.

GPRs 2 and 3 have been explicitly tested by asking listeners to identify the location of boundaries (e.g., Clarke & Krumhansl, 1990; Deliège, 1987; Peretz, 1989, Experiment 1). In addition, numerous studies related to parsing have interpreted results on the basis of structures that resemble GPRs 2 and 3 (e.g., Boltz, 1989, 1991; Deliège, 1989; Deliège & El Ahmadi, 1989; Deliège, Mélen, Stammers, & Cross, 1996; Dowling, 1973; Gregory, 1978; Jusczyk & Krumhansl, 1993; Sloboda & Gregory, 1980; Stoffer, 1985; Tan, Aiello, & Bever, 1981). Indirect tests have been provided by studies that implicitly use structures based on GPRs 2 and 3 (e.g., Krumhansl, 1996; Palmer & Krumhansl, 1987; Peretz, 1989,

Experiments 2 and 3; Peretz & Babai, 1992). While supporting the conceptualization of the *GTTM*, close examination reveals several methodological and theoretical concerns. The present work builds upon this previous research by quantifying the GPRs *as defined within GTTM* (cf. Lerdahl & Jackendoff, 1983) while implementing several methodological improvements.

The Need for Quantification

Quantification of each rule (or equivalently, a precise operational definition) is necessary for a proper test of any theory. Even if the theory is based on intuitions (*GTTM* is not an axiomatic theory), quantification makes those intuitions explicit and forces them to be applied consistently. With respect to *GTTM*, quantification will automatically detect every instance of every rule within a given stimulus while permitting a continuous rather than binary (yes/no) coding of rule application. This facilitates comparison of (1) applications of the same rule at different points in one or more stimuli, (2) different rules at the same point in a single stimulus, and (3) different rules at different points in one or more stimuli (possibly in different experiments). Without quantification, these opportunities are reduced.

For example, Peretz (1989, Experiment 1) tested Register change (GPR 3a), Length change (GPR 3d), and Parallelism (GPR 6) using nine short monophonic extracts (<15 notes). Each extract contained at least one boundary based on these rules, but only one particular boundary per sequence was analyzed. Participants were asked to identify the natural breaks in the melody by reference to a line of dots that matched the notes. Both musicians and nonmusicians performed in high concordance with the rules (87.5% and 77.1%, respectively). Why was the concordance not 100%? Although the details of "errors" were not discussed, it is possible that other potential boundaries were stronger. Simple visual inspection of the 9 stimuli provides 11 *additional* possible applications of Register change, 1 *additional* application of Length change, and 6 possible applications of Attack-point (GPR 2b; the stimuli presented as Length could have represented Attack-point). If participants did parse on these alternative locations, it was not noted. If participants did not parse on these alternative locations, then what were the errors? Quantification would identify all applications of all GPRs, providing an assessment of the relative importance of each and a more thorough test of the theory.

In a similar work, Deliège (1987, Experiment 1) presented participants with 32 short extracts (3 to 16 notes) from the Western-European repertoire and asked them to indicate parsing with reference to a line of dots

that matched the number of sounds in the upper voice. As with the work of Peretz (1989), both musician and nonmusician groups parsed in accordance with the GPRs, and musicians did so significantly more often. Deliège found that the different rules had different utilities: For both groups, Attack-point (GPR 2b) had the highest and Articulation change (GPR 3c) had the lowest. However, because the stimuli were selected for the GPRs that they contained, selection may have produced strong versions of Attack-point and weak versions of Articulation change. Quantification would do much to alleviate such concerns. Furthermore, the stimuli may have contained more than one rule. For example, Deliège (1987, p. 343) noted that one stimulus (Beethoven's String Quartet, shown in her Figure 14), intended to test parsing on the basis of Length change (GPR 3d), could have represented parsing on the basis of a "change of instrumental and/or sound density." Quantification would have identified the potential contributions of all rules, thereby helping to delineate their effects. Implying a need to quantify, Deliège split Register change (GPR 3a) into two rules (small and large changes).

Deliège (1987, Experiment 2) presented participants with sequences designed to contain two competing GPRs in adjacent locations. Participants were asked to provide one parsing of each sequence, thereby demonstrating the stronger rule. The results were used to develop a hierarchy of rule strength. For musicians, Dynamic change (GPR 3b) was the strongest and Length change (GPR 3d) the weakest. For nonmusicians, Register change (GPR 3a) was the strongest and Length change the weakest. However, the resulting hierarchy may be an artifact of stimuli that compared strong versions of Dynamic change or Register change with weak versions of Length change. Quantification would help to alleviate such concerns. As noted by Clarke and Krumhansl (1990, p. 216), "It is only sensible to consider the relative strength and/or weakness of different rules if some kind of quantitative comparison can be made . . . At present, no such interparametric metric exists."

Clarke and Krumhansl (1990, Experiments 1 and 4) presented highly trained participants with three repetitions of two pieces for piano by Stockhausen (Experiment 1) and Mozart (Experiment 4). While listening to the second presentation, participants indicated boundaries for "relatively large-scale segments" (p. 225) by pressing a foot pedal. On the third presentation, each piece was interrupted at the previously identified boundaries, and participants indicated the exact location of the boundary by reference to the score, the strength of that boundary (7-point scale), the ease of boundary localization (7-point scale), and the musical features that had "caused" the boundary. Multiple causes were cited for the boundaries identified, and many of these corresponded to the GPRs of *GTTM*. However, with-

out quantification of the GPRs, the empirical data cannot be used to validate the theory.

Quantification of GPRs 2 and 3 (and GPR 4)

Given this demonstrated need for quantification, what follows is the quantification of GPRs 2 and 3 of *GTTM*. The four-note span that defines each GPR (see Table 1) specifies the location of a boundary but not its strength (however, no boundary must imply a strength of 0). *GTTM* implies that boundary strength should relate to the degree of rule adherence (e.g., larger intervallic distances should produce stronger boundaries). Proper quantification also requires careful specification of the appropriate basis and the scaling for each rule (these are intertwined). Improper quantification might hide important relationships or emphasize unimportant ones, tantamount to testing a theory other than *GTTM*.

The Rest aspect of Slur/Rest (GPR 2a), Attack-point (GPR 2b), and Length change (GPR 3d) concern perceived duration. As such, in the present quantification scheme, all were based on linearly scaled time, consistent with psychoacoustic demonstrations of a linear scaling for the perception of time intervals with durations in the range of musical notes (for reviews, see Allen, 1979; Handel, 1993). Hence, a quarter note/rest was assumed to be twice as long as an eighth note/rest and one-half the length of a half note/rest.

Slur/Rest (GPR 2a), as defined within *GTTM*, subsumes rests and slurs within the same rule. The process of quantification leads to the realization that combining rests and slurs within one rule is not ideal. A slur is an aspect of the internote interval (the interstimulus interval or ISI). This interval contributes to the articulation of notes (musical events) as slurred, legato, or staccato. A rest is the absence of sound at a point where a note could potentially occur, with a duration comparable to a note. Even though a rest does not have an acoustic onset per se, a listener will know that a rest has occurred because expectancies arise from the prior temporal pattern of musical events. That is, the meter of a piece informs listeners of *when* to attend to musical events (cf. Jones & Boltz, 1989). A hierarchical structure of meter (cf. Lerdahl & Jackendoff, 1983) could inform listeners of the relative importance of individual events. As such, the perceptual meaning of a rest is not necessarily comparable to the perceptual meaning of a slur (or ISI). In this work, only the rest aspect of Slur/Rest (hereafter *Rest*) was quantified as the absolute magnitude of the rest. A whole-note rest was coded as a boundary potential of 1.0, with other rest values being scaled proportionally, so that boundary strength ranges from 1/64 to 1.0. A 64th note is the smallest temporal value specifiable in standard musical notation. Rests longer than a whole note should be assigned

a value of 1.0 (rests longer than a whole note are rare in monophonic music). The location of the rest defines the location of the boundary.

Attack-point (GPR 2b) implies that relatively longer time differences between note onsets create stronger boundaries. The simplest implementation compares the attack-point interval between notes n_2 and n_3 with the average of the intervals between n_1 and n_2, and that between n_3 and n_4. This reduces to the length of n_2 compared with the average of n_1 and n_3:

$$\text{boundary strength} = 1.0 - \frac{n_1 + n_3}{2 \times n_2}, \text{ where the n's are lengths.} \qquad (1)$$

For the rule to apply, n_2 must be longer than both n_1 and n_3. In addition, n_1 through n_4 must be notes. These conditions are necessary to separate Attack-point from Rest (GPR 2a) and Length change (GPR 3d). It is assumed that by defining different GPRs, *GTTM* intends those GPRs to address different effects. Attack-point produces a boundary value of 0.0 when $n_2 = n_1 = n_3$ and a value near 1.0 if the lengths of n_1 and n_3 are much smaller than that of n_2 (e.g., for a whole note surrounded by 32nd notes, the value is 0.97). Quantification as a ratio removes consideration of the units (e.g., milliseconds) used to measure note duration. Figures 1 through 6 (Panel A) provide some examples of the application of this rule in six different melodies.

For Length change (GPR 3d, hereafter Length), the length of n_1 must equal n_2, and the length of n_3 must equal n_4, so the simplest quantification is:

$$\text{boundary strength} = 1.0 - \frac{n_1}{n_3} \quad \text{if } n_3 > n_1, \text{ where the n's are lengths.}$$

or (2)

$$\text{boundary strength} = 1.0 - \frac{n_3}{n_1} \quad \text{if } n_3 < n_1, \text{ where the n's are lengths.}$$

This coding results in a value of 0.0 when the two pairs have the same length and a value near 1.0 when the lengths are very different (e.g., for whole notes compared with 32nd notes, the value is 0.97). Figures 1, 3, 5, and 6 provide some examples. Note that for Length to apply, one must wait until n_4 has finished in order to assess the length of n_4.

Register change (GPR 3a; hereafter Register) concerns the perception of pitch. Register, as defined within *GTTM*, strongly implies chromatically scaled pitch heights (logarithmically scaled frequencies: equal-tempered tuning) by a lack of reference to actual frequencies or to notions of tonality. Obviously, the scaling of pitch in music must be at least chromatic (cf. Cohen, Trehub, & Thorpe, 1989; Cuddy & Cohen, 1976; Krumhansl,

1979; Krumhansl & Shepard, 1979; Shepard, 1982), but should the scaling be diatonic? Even if a diatonic scaling is intended, *GTTM* (see also Jackendoff, 1992; Lerdahl, 1988a, 1988b, 1992) does not provide a method for quantifying Register within a particular tonal framework (i.e., the relationships between adjacent steps within a particular key). For example, if the key is C major, then a change from C to D or E to F could be defined as one unit, but then, would a change from C to C♯ be one-half unit (even though, like E to F, it is 1 semitone)? Furthermore, if Register is referenced to a particular scale or key, then the method by which listeners abstract the most appropriate scale or key, *while listening,* must be defined (cf. Cohen, 1991, 2000; Frankland & Cohen, 1996; Vos & Van Geenen, 1996). There is no consensus in the literature, and *GTTM* does not define a method. Lerdahl (2001) provides some interesting extensions to *GTTM* but does not yet provide a quantifiable method for assessing tonality or register. For these reasons, Register was based on the simpler chromatic scale. Register also requires consideration of absolute versus relative magnitudes of change. Because relative change seems to be implied by the definition of phenomenal accents (Lerdahl & Jackendoff, 1983, p. 17), Register was quantified as:

$$\text{boundary strength} = 1.0 - \frac{|n_1 - n_2| + |n_3 - n_4|}{2 \times |n_2 - n_3|} \quad \text{where the n's are pitch heights in MIDI notation} \qquad (3)$$

The rule applies only if the transition from n_2 to n_3 is greater than from n_1 to n_2 and from n_3 to n_4. In addition, the transition from n_2 to n_3 must be nonzero. To equate rising and falling pitch contours, absolute values were used. Pitch height was expressed using the standard Musical Instrument Device Interface (MIDI) format (i.e., A_4 = 440 Hz = note 57). The use of a ratio of relative change produces a boundary value of 0.0 when there is no difference in the size of the intervals and a value near 1.0 when the interval distance from n_2 to n_3 is much larger than the average interval distance from n_1 to n_2 and from n_3 to n_4. For example, for the sequence E_4-F_4-B_4-C_5, the value is 0.83, and for the sequence C_4-D_4-F_4-G_4, the value is 0.33. Figures 1 through 6 provide some examples in different melodies. Note that any relative scaling will assign a large boundary value to a small interval change in an otherwise flat melodic contour (e.g., C_4-C_4-D_4-D_4). To avoid this problem, a quantification based on the absolute magnitude of each interval could be used. However, such a quantification tends to assign too much importance to a large change in the midst of other large changes (e.g., D_4-G_4-C_4-F_4). In the current work, Register was quantified by using both relative and absolute scalings, but the relative scaling performed better so only the relative scaling is presented here.

Thus, for the present work, four specific rules (Rest, Attack-point, Length, and Register) were quantified. To facilitate comparisons between rules, all rules were scaled to the range from 0.0 to 1.0, with 0.0 implying no boundary and 1.0 implying the strongest possible boundary for that rule. At this time, the strengths of different rules can be directly compared only at 0.0 (no strength means no boundary): A rule strength of 1.0 does not necessarily mean a 100% chance of a boundary—the strongest possible versions of some rules may not always induce a boundary. Similarly, a rule strength of 0.5 may represent an 80% chance of a boundary for one rule and a 20% chance for a different rule. The actual relationship between rule strength and boundary formation is an empirical question, to be answered, in part, by this work. In addition, in this quantification, each rule was represented by a linear function (in the range of 0 to 1, linear and nonlinear functions tend to be highly correlated). Finally, in the application of all rules, tied notes were treated as a single note with the same total duration and all internote durations (ISI) were considered as 0.0 (eliminating any articulatory delineation of notes as slurred, legato, or staccato).

Although quantification focused on GPRs 2 and 3, it is possible to include Intensification (GPR 4; see Table 1). As stated, Intensification implies that the probability of a boundary should be monotonically related to the rule strength. This aspect of Intensification was encoded directly into each GPR separately. However, Intensification also implies (but does not explicitly state) that when two or more aspects of GPRs 2 and 3 coincide, there should be a higher probability for a boundary. As such, Intensification could represent the combined action of GPRs 2 and 3, which can be quantified using multiple regression.

Methodological Enhancements

When choosing or designing stimuli to test GPRs, it must be remembered that *GTTM* claims relevance only to Western tonal music and listeners experienced in that idiom (cf. Lerdahl & Jackendoff, 1983, p. 4, pp. 36–42). As such, all stimuli should, as a minimum, evoke the same kind of processing that listeners would use when processing such music. On the other hand, GPRs 2 and 3 can be tested only with monophonic stimuli because neither *GTTM* nor its extensions (Jackendoff, 1992; Lerdahl, 1988a, 1988b, 1992, 2001) is sufficiently well developed for quantified predictions to be applied in polyphonic, or even homophonic music. The GPRs (see Table 1) refer only to monophonic, four-note spans. It is unclear how GPRs 1 through 6 can be extended to complex homophonic music: Only one example of the application of the GPRs within a complex

piece (p. 66) was provided. *GTTM* assumes that "a single grouping analysis suffices for all voices of a piece" (Lerdahl & Jackendoff, 1983, p. 37), but it is obvious that, except in highly constrained situations, the application of the GPRs 2 and 3 is not the same for all voices (cf. Temperley, 2001, p. 63), presupposing one knew how listeners delineated the melody from its accompaniment or the different voices from each other (cf. Acker & Pastore, 1996; Bozzi, Caramelli & Zecchinelli, 1994).

Hence, in this work, stimuli were restricted to monophonic melodies from Western tonal music, consistent with the approach of Peretz (1989). In addition, when designing stimuli, the ideal of strong internal validity requires that only one aspect of the stimulus be changed at a time, but for music, strong internal validity tends to conflict with external validity (i.e., the use of stimuli representative of Western tonal music). To balance such concerns, it was decided, a priori, to control aspects of the stimuli that could be controlled without an unacceptable loss of musicality. Therefore, in the present experiments, the stimuli did not contain any information that could be used to generate boundaries on the basis of the slur aspect of Slur/Rest (GPR 2a), Dynamic change (GPR 3b), or Articulation change (GPR 3c). All tones were presented at the same intensity, and all ISIs were 0.0. For that reason, these rules have not been presented.

To obtain empirical (subjective) boundary locations on a note-by-note basis, the online task of Clarke and Krumhansl (1990), Deliège (1989), Deliège and El Ahmadi (1989), and Krumhansl (1996) is preferable to the offline task of Deliège (1987) and Peretz (1989). In the online method, participants parsed short sequences of music, while listening to the music, using a simple key press. In the offline method, participants placed a marker on a line of undifferentiated dots, *after listening* to the entire piece. For the assessment of GPRs 2 and 3, the online method seems less likely to be confounded by retrospective reinterpretations of the boundary locations based on global considerations of the melodic structure (i.e., GPR 7). In addition, the offline method is limited to short sequences of notes because a participant would be likely to lose serial position in a longer work (cf. Deliège, 1987, p. 335). Although the inclusion of place markers in the line of dots might help, such place markers might induce parsing biases. Similarly, using a representation of the score (cf. Clarke & Krumhansl, 1990; Deliège, 1989; Deliège & El Ahmadi, 1989) might induce parsing on the basis of the visual-spatial pattern rather than the auditory-temporal pattern (cf. Cook, 1989, p. 119). On the other hand, in the online task, there are concerns about the reaction times of participants. While listening, participants must detect a boundary (create a unit) and then respond with a key press. Although reaction time analysis is notoriously complex (see Luce, 1986), it is reasonable to assume that the fastest auditory detection reaction times are

on the order of 100 ms. Typically, decision reaction times are much longer (perhaps 250 to 300 ms) and more variable. These values fall within the range of durations of notes in typical music, so reaction times are an issue, particularly since previous studies have not obtained fine temporal resolution.

In the work of Deliège (1989) and Deliège and El Ahmadi (1989), a tape playback machine presented the music while, in parallel, a computer recorded the responses of participants (as time elapsed from start). The start of response recording was synchronized to the start of the music. This cumbersome method required a manual regrouping of key press responses to capture the most reasonable boundary location (relative to the score). As such, there was a loss in resolution and a danger of coder bias (who decides what is "reasonable"?). Clarke and Krumhansl (1990) used a foot pedal to indicate boundaries but found that this "key press" could provide only an approximate boundary location. Participants had to indicate the exact location of the boundary by reference to the score. Krumhansl (1996, p. 409) used a mouse click to indicate boundaries but found it necessary to average the responses of participants over a two-beat window in order to "capture the clustering of responses," implying low temporal resolution. On the other hand, a series of studies in the area of social psychology, using a key press to parse continuous visual sequences of events, has demonstrated that participants can produce reliable and meaningful parsings with fine temporal resolution for behavioral sequences as short as 6 s and as long as 7 min (e.g., Newtson 1973, 1976; Newtson, Engquist, & Bois, 1977). The act of parsing (key pressing) did not interfere with the normal perception of events, and participants found the procedure simple to learn. Detailed analyses of consistency demonstrated within-subjects reproducibility across repetitions and between-subjects similarity, both of which one might expect in music. This more rigorous approach was adapted for the empirical test of the parsing rules of *GTTM* in the following two experiments.

Experiment 1

In Experiment 1, while listening to each of three melodies, participants used a key press to indicate the end of one unit and the beginning of the next (i.e., the boundary between units). They were requested to make their units as small (but meaningful) as possible. For each participant, for each melody tested, a *boundary profile* was created.

All stimuli were relatively simple monophonic melodies, chosen quasi-randomly from collections that were arranged for elementary instruction in piano. As such, stimuli were *neither designed nor selected* for the rules

that they contained. Those rules that happened to be in those melodies were tested. Some melodies were selected to be simple and familiar (hence, predictable parsings, possibly linked to the lyrics), and some were chosen to be unfamiliar (to prevent parsing based on lyrics). All melodies were presented at constant intensity, and all articulation was legato. This eliminated parsing on the basis of Slur (GPR 2a: Slur), Dynamic change (GPR 3b), and Articulation change (GPR 3c), while minimizing parsing on the basis of metrical structure (beats discerned from changes in intensity or small changes in relative timing). Thus, only Rest (GPR 2a: Rest), Attack-point (GPR 2b), Register (GPR 3a), Length (GPR 3d), and their combination (Intensification: GPR 4) could be used to parse the melodies.

There were three main analyses. Within-subject reliability of boundary placement (labeled *consistency analysis*) was assessed by the use of repeated trials for each melody and a subsequent correlational analysis of the pattern of boundaries across repetitions. High consistencies were neither required nor necessarily expected. Low consistencies would provide evidence of learning and/or the effects of top-down processing (i.e., GPR 7). In the second analysis (labeled *similarity analysis*), between-subject reliability of boundary placement was assessed through correlational and cluster analyses. Previous studies have demonstrated a high degree of similarity, but there have been relatively minor differences related to training or experience (cf. Deliège, 1987, 1989; Krumhansl, 1996; Peretz, 1989). Hence, the expectation was that all participants would form one homogeneous group. The final analysis related the boundary profiles to the quantified GPRs of *GTTM*. Comparisons between each GPR and the empirical data were essentially correlational analyses. Because no one rule could be expected to capture all the empirically determined boundary positions, multiple regression was used to test the combination of all rules (Intensification: GPR 4). In these analyses, the GPRs were compared to the grouped empirical data (from the similarity analysis). As such, it was assumed that those places where many participants placed a boundary corresponded to a strong boundary and vice versa. Additional analyses examined individual differences.

It must be noted that the work presented here is only a small, but necessarily first, part of a much larger project that examined the relationship between training, tonality, other indices of musical involvement and parsing, while extending the analysis to Parallelism (GPR 6). As such, several aspects of design were constrained by other aspects of the project. Discussion of those aspects is only included where it provides necessary context. To that end, some discussion of the two experimental contexts for the boundary placement (boundary-efficacy tasks) is necessary for the proper interpretation of results.

THE BOUNDARY-EFFICACY TASKS

For any melodic parsing task, participants may create reliable parsings; however, those parsings may simply reflect the demand characteristics of the experiment without any real relevance to the processing or storage of the music. Two different experimental contexts for melodic parsing were developed to validate the parsing of each melody (cf. Newtson, 1973, 1976; Peretz, 1989). The two tasks are referred to as *boundary-efficacy tasks.* One group of participants was assigned to each.

After first listening to and parsing the melody, participants in Group 1 were presented with a recognition-memory task, modeled after Peretz (1989). Sixteen four-note probes were extracted from the melody, and 8 of the 16 probes were altered by changing one note of the probe by ±1 or ±2 semitones. Participants were presented with each probe and indicated whether or not each had been in the melody that they had just parsed: Responses were a binary yes/no. Critically, probes were selected such that eight straddled the boundaries that the participant had previously identified, and eight did not. Following Dowling (1973), Peretz (1989), Peretz and Babai (1992), and Tan et al. (1981), one would predict lower recognition performance for probes that straddled boundaries. The results indicated that the boundaries identified by the participants had some validity for the storage of the melody.

After listening to and parsing the melody, participants in Group 2 were presented with a click-detection task, modeled after Berent and Perfetti (1993), Gregory (1978), Sloboda and Gregory (1980), and Stoffer (1985, Experiment 2; see also Clark & Clark, 1977; Kahneman, 1973, cited in Berent & Perfetti). Participants were asked to detect the presence of clicks embedded in the melodies, while listening to the melodies. In this task, it is assumed that participants must split their processing resources between the two tasks: listening to the music and responding to clicks. When the music demands more processing resources, fewer processing resources are available to the click-detection task. If boundaries represent the closure of a unit (hence, additional processing), then in the moments prior to closure, the music demands more processing. As such, reaction times to detect a click on, or just in front of, a boundary would be longer than reaction times to detect a click in the middle of a unit. Eight clicks were placed within the melody online, based on the boundaries that the participant had previously identified in the melody. Four clicks were placed on randomly selected boundaries, and four clicks were placed randomly between boundaries. Again, the results indicated that the boundaries identified by the participants had some validity.

Detailed results of the boundary-efficacy tasks are not presented here for reasons of brevity. The important point is that the recognition-memory task of Group 1 requires memorization. One cannot compare the probe

to the melody unless one has, in some sense, remembered the melody. Conversely, the click-detection task of Group 2 did not require memorization. Participants could have responded, as easily, to the presence of clicks during the first presentation of the melody. If the two groups produced the same, interpretable, boundary profile, then memorization of the melody did not affect parsing. If only Group 1 produced an interpretable boundary profile, then higher processing is required for meaningful boundary formation. If only Group 2 produced an interpretable boundary profile, then the demands of higher processing interfere with meaningful boundary formation. Finally, if the two experiments produced different, but interpretable, boundary profiles, then higher processing forces a different kind of parsing from simple listening. In any event, a dissimilarity in the profiles for the two experiments would be cause for further investigation. Since no major differences emerged in the patterns of parsing between the two groups, the results for the two groups are presented together.

METHOD

Participants

In Experiment 1, there were 123 participants (80 females) with a mean age of 22.09 (*SD* = 6.35) years (range = 16–52 years), with the equivalent of an average of 11.7 (*SD* = 18.9) years of instruction at 1 hr per week. All participants were recruited from the university community, primarily the departments of psychology and music. Royal Conservatory of Music (RCM) or equivalent grades ranged from 0 to beyond the Associateship level of Grade 11.

Procedure

Each experiment had eight stages. Stages 1, 3, 5, and 7 assessed the internal representation of tonality of the participant using a modified probe-tone task (cf. Frankland & Cohen, 1990; Krumhansl, 1990). As noted previously, this article is focused on the parsing of melodies, so Stages 1, 3, 5, and 7 will not be discussed further. Stages 2, 4, 6, and 8 assessed boundary formation within particular melodies and determined the efficacy of those boundaries. Stages 2, 4, 6, and 8 differed only in the melodic stimuli. Detailed instructions were presented by computer to the participant at the beginning of each stage; abbreviated instructions remained on screen during each stage. The experimenter remained with the participant during practice (Stages 1 and 2) to ensure, by direct observation, that the instructions were understood and to provide further clarification if needed. Participants were seated comfortably in front of a computer in a sound-attenuated room for the entire experiment.

Stages 2, 4, 6, and 8, each had two parts. The first part was designed to assess the location of boundaries within a melody. Participants listened to the melody and simultaneously indicated the location of boundaries by a key press. The second part (not presented here) assessed the efficacy of those boundaries as they pertain to the formation of informational units, using the recognition-memory task (Group 1, $n = 61$) or a click-detection task (Group 2, $n = 62$).

In Part 1 of each of Stages 2, 4, 6, and 8, a melody was presented. Participants were instructed to press the space bar (or any key of their choosing) any time they felt that one section of the melody had ended and a new section of the melody had begun: The term used was *break*. The analogy with the parsing of a line of speech into its component words

was stated explicitly. Participants were informed that there would be three trials using the same melody, and that each subsequent repetition was intended to allow them to refine their answers. In Stage 2 (Practice), additional verbal instructions were provided if needed.

At the end of the first presentation of the melody, participants were asked to rate the familiarity of the melody on a continuous scale from 0 (*unfamiliar*) to 10 (*familiar*). After rating the familiarity, participants were asked to provide the name of the melody, or a line from the lyrics. Participants then heard the melody and indicated boundaries for the second and third repetitions. Participants initiated each presentation of the melody at the time of their own choosing, but the timing within the melody was fixed for all participants.

After the third repetition, participants moved on to Part 2, again at the time of their choosing. However, participants were informed, at the beginning of each stage, that there would be a recognition-memory task (Group 1) or click-detection task (Group 2) to follow the final presentation of the melody. The reminder was to discourage long delays between the parsing and memory task of Group 1, but it was retained for consistency for Group 2.

Upon conclusion of Stage 8, participants completed the questionnaire pertaining to musical background and were debriefed. The entire experiment lasted no more than 1 hr.

Apparatus and Stimuli

Tones were created using the default instrument 0, mode 0, (an acoustic piano sound) of the internal MIDI driver of a Creative Labs Sound Blaster 16, housed within an IBM AT (80286, 12 MHz) compatible computer. The same computer provided instruction via a B/W monitor and recorded responses. Programs for stimulus presentation with associated response collection and for quantification were written in-house (by B.W.F.), using Borland's Turbo C/C++, Version 3.0, aided by the Creative Lab's Sound Blaster Developer's Kit, Version 2.0. The notes of all melodies were presented at the same intensity with precise computer-controlled timing.

Tones were presented binaurally through Realistic LV 10 headphones connected directly to the audio output of the Sound Blaster at a level considered comfortable by the participant. The monitor and keyboard were housed in the Industrial Acoustics single-walled, sound-attenuating room, but the main computer was external to this room to minimize noise.

In Stage 2, the melody presented to participants was "The Mulberry Bush" (see Figure 1) with notes in the range G_4 to G_5 (all notes are referenced to A_4 = 440 Hz) played at a tempo of 135 beats per min (quarter note = 444 ms). When assessed by the Krumhansl and Schmuckler key-finding algorithm (Krumhansl, 1990), the melody, as a whole, had a key strength of r^2 = .69 for the best-fitting key C major) and r^2 = .41 for the second best key C minor) yielding a *q*-factor of .28 (*q* ranges from 0 to 1, with 1 meaning an unambiguous key; see Frankland & Cohen, 1996 for more a detailed discussion). This melody is tonal. This simple nursery rhyme had a moderate familiarity, achieving a mean rating of 5.99 (*SD* = 3.14, median = 7.00, mode = 10, range = 0–10). Six participants correctly labeled the tune, four by name and two by citing the first line of the melody. In subsequent discussions, it will be referred to as "Mulberry."

Stage 4 presented the melody "Three Blind Mice" (see Figure 2) with notes ranging from D_4 to D_5. The tempo was set to 220 beats per minute (quarter note = 273 ms). The computed key strength was r^2 = .90 for the best fitting key (D major) and r^2 = .41 for the second best key (F♯ minor), producing q = .49. This melody is also unambiguous in its tonal center. This nursery rhyme had a high familiarity generating a mean familiarity rating of 9.07 (*SD* = 2.30, median = 10.00, mode = 10, range = 0–10). A total of 99 participants correctly identified the melody, while 7 participants mislabeled it with the name of a different nursery rhyme. In subsequent discussions, it will be referred to as "Mice."

In Stages 6 and 8, the melody presented was "Softly Now the Light of Day" (see Figure 3; the same extract was used by Boltz, 1989) with notes ranging from B_4 to B_5, at a tempo of 150 beats/min (quarter note = 400 ms). The computed key strength was r^2 = .51 for the

best-fitting key C major) and $r^2 = .43$ for the second best key (E minor), producing $q = .07$. Although tonal, it is more ambiguous with respect to its tonal center than the previous two (i.e., lower key strength and lower q factor). The same melody was used in Stages 6 and 8 because, a priori, it was felt that the participants might require more repetitions to stabilize their responses within an unfamiliar melody (i.e., the consistency analysis). At the beginning of Stage 6, this melody was very unfamiliar to participants in achieving a mean rating of 1.57 (SD = 2.06, median = 1.00, mode = 0, range = 0–10). No one identified this tune. In subsequent discussions, it is labeled "Softly."

During the collection of responses, key presses were linked to the currently sounding note at the time of the key press. The boundary must have occurred before the key press. As such, a participant could have used the key press to indicate the end of a functional unit (i.e., the last note of a unit), or the beginning of a functional unit (i.e., the first note of the next unit), or the boundary between units. It is extremely unlikely that any participant could have timed key presses with sufficient precision to indicate the actual boundary between units. The boundary between units corresponds to the gap between the offset of the last note of the prior unit and the onset of the first note of the subsequent unit. The gap between notes is the time that the CPU takes to instruct the soundboard to turn off one note and to turn on the next note. Although this gap would depend on instrumentation, in this work it was only a few *microseconds*, while the fastest human reaction times are on the order of 100 *milliseconds*. Hence, key presses of the participant could indicate only either the last note or the first note of a functional unit. To determine which, the responses of each participant were statistically compared to the predictable parsings for the simple nursery rhymes of Stages 2 and 4; hence the need for, and value of, these simple melodies with highly predictable parsings. By aligning the participant's boundaries with this predicted pattern of boundaries, it was possible to determine whether a participant placed the boundary after the end of a unit, or before the beginning of a unit. This analysis indicated that participants universally used their key presses to indicate the last note of a unit. It was assumed that the same relationship held for the melody of Stages 6 and 8. For the purpose of presentation, the boundaries identified by participants have been placed between units.

RESULTS AND DISCUSSION

The main focus of the boundary analysis was the assessment of the placement of boundaries and the assessment of the relationship between empirically determined boundary locations and the quantified model of *GTTM*. Several other analyses were conducted to ensure the integrity of any conclusion based on these analyses. Each participant parsed "Mulberry" (Stage 2) three times, "Mice" (Stage 4) three times, and "Softly" six times (three in each of Stages 6 and 8), producing a single *empirical boundary profile* for each repetition. For each participant, for each stage, for each repetition, the empirical boundary profile was converted to a binary coding of the presence of boundaries between successive notes. A value of 1 indicated a boundary between adjacent notes, and a value of 0 indicated no boundary. For each stage, the length of the boundary profile depended on the number of notes in the melody, with $N = 35$ for "Mulberry," $N = 47$ for "Mice," and $N = 33$ for "Softly" (hereafter, an uppercase N denotes the melody length while a lowercase n denotes number of subjects).

For all analyses, there were $n = 123$ participants, divided into two groups (i.e., the different boundary efficacy tasks). In the consistency and

similarity analyses, the responses of Groups 1 and 2 were explicitly compared. For reasons to follow, the two groups were collapsed for all subsequent analyses. Though considered practice, the analyses of Stage 2 have been included because their results were congruent with those of the other stages.

Consistency Analysis

The consistency analysis examined the relationship between the boundary profiles for each repetition, for each participant individually. Ideally, high consistencies would permit the averaging of responses across repetitions. For each melody, for each participant separately, correlations were computed between the boundary profiles produced on the different repetitions. The phi coefficient[1] was used to compute these correlations because it is the binary analogue of Pearson's correlation (i.e., the usual r). To provide some intuition for the meaning of these values, assume that a single participant had parsed the same melody on two repetitions. If that participant placed boundaries at the same locations only 50% of the time, then the correlation would be $r = .43$ assuming eight notes per unit, $r = .40$ assuming six notes per unit, or $r = .33$ assuming four notes per unit. The actual significance of these correlations is determined by their magnitude and the number of notes in the boundary profile (N).

The consistencies between repetitions for each melody are shown in Table 2, along with the average consistency per stage computed as the mean of means. In Table 2, for reasons of space, the nine correlations between the repetitions of Stage 6 and those of Stage 8 are collapsed into a single average.

For each melody, the consistencies were relatively high. All distributions were negatively skewed. The modal correlations were 1.0, except those involving Repetition 1 of "Mulberry" and the comparison of Repetitions 1 and 3 of "Softly." For all melodies, the mean correlations were significantly different from $r = 0$ ($p < .001$) using a one-group t-test, implying that the different repetitions could be averaged. Note that Repetitions 2 and 3 of each stage produced higher mean correlations than 1 and 2, or 1 and 3.

When examining each participant individually, any consistency of $r \geq .33$ would be significant ($p < .05$) for "Mulberry" (Stage 2; $N = 35$). Similarly, any consistency of $r \geq .30$ would be significant for "Mice" (Stage 4, $N = 47$). The corresponding value is $r \geq .34$ for "Softly" (Stages

1. Boundary profiles were analyzed by using the simple similarity measure, the Jaccard similarity measure (or similarity ratio), the Kulczynski 2 similarity measure, the Sokal and Sneath 4 similarity measure, and the phi similarity measure (cf. Howell, 2002; SPSS, 1988). All measures were highly correlated using both Spearman's rank and Pearson's statistics ($r \geq .97$) across different participants so only the phi is reported.

TABLE 2
The Average Consistencies Within Subjects for "Mulberry" (Stage 2), "Mice" (Stage 4), and "Softly" (Stages 6 and 8) in Experiment 1

Melody	Repetitions	*M*	*SD*	Min	# Sig[a]	L-M[b]
Mulberry	1 to 2	.527	.289	–.187	95	69
	1 to 3	.519	.311	–.294	90	67
	2 to 3	.764	.227	–.139	114	110
	Stage 2[c]	.603	.139		88[d]	61[d]
Mice	1 to 2	.703	.223	–.110	118	110
	1 to 3	.674	.251	–.110	114	105
	2 to 3	.733	.249	–.045	115	111
	Stage 4[c]	.703	.030		111[d]	98[d]
Softly	1 to 2	.417	.333	–.138	70	42
	1 to 3	.386	.310	–.179	67	36
	2 to 3	.549	.310	–.157	90	62
	Stage 6[c]	.451	.087		47[d]	24[d]
	Between[e]	.500	.064			
	4 to 5	.545	.310	–.318	93	64
	4 to 6	.529	.335	–.149	87	56
	5 to 6	.590	.340	–.175	93	65
	Stage 8[c]	.555	.032		72[d]	43[d]

NOTE—Maximums were uniformly 1.000, so they are not included.
[a]# Sig is the number of participants out of 123, who produced a significant correlation ($p < .05$) between repetitions.
[b]L-M is the Larzelere and Muliak test that applies a Bonferroni correction to # Sig.
[c]Stage provides the means and standard error of each stage.
[d]Number of participants who had significant correlations between all three repetitions in a stage.
[e]Between provides the mean and standard error of the 9 correlations comparing Stage 6 to 8.

6 and 8; $N = 33$). Using these values, Table 2 also presents the number of participants who had significant correlations between repetitions within each stage. To protect Type 1 error rate, the Larzelere and Muliak test (Howell, 2002) was also used. Note that, for all stages, between 73% and 97% of all participants had a significant correlation between Repetitions 2 and 3 (50% and 90% for the Larzelere and Muliak test).

The consistencies between repetitions for individual participants were compared using a mixed analysis of variance (ANOVA) with Group as the between-subjects factor and Repetition Consistency as the within-subjects factor. There were three Repetition Consistencies (i.e., correlations) for "Mulberry" (Stage 2) and "Mice" (Stage 4), but there were 15 Repetition Consistencies for "Softly" (Stages 6 and 8). For the purpose of these analyses, the Fisher *r* to *z* transformation was applied to each individual correlation. There was no effect of Group for "Mulberry," $F(1, 121) = 1.46$, *ns*, "Mice," $F(1, 121) = 1.94$, *ns*, or "Softly," $F(1, 121) = .15$, *ns*. Effect sizes (η^2 and ω^2) were always less than 0.01. There were effects of Repetition Consistency for "Mulberry," $F(2, 242) = 36.09$, $p < .001$, $\eta^2 = .228$, $\omega^2 = .221$, "Mice," $F(2, 242) = 5.58$, $p < .004$, $\eta^2 = .044$, $\omega^2 = .036$,

and "Softly," $F(14, 1694) = 9.12$, $p < .001$, $\eta^2 = .069$, $\omega^2 = .062$. For "Mulberry" and "Mice," orthogonal contrasts using a Bonferroni correction ($p < .025$) demonstrated that the correlation between Repetitions 2 and 3 was higher than the average correlation between Repetitions 1 and 2 and Repetitions 1 and 3. For "Softly," the 14 orthogonal contrasts using a Bonferroni correction ($p < .0036$) indicated that the correlation between Repetitions 2 and 3 of Stage 6 was higher than the others in Stage 6 and that the correlation between Repetitions 5 and 6 was higher than the others in Stage 8. However there were no differences in the correlations for any combination of Repetitions 2, 3, 5, and 6. There were no interactions between Group and Repetition Consistency for "Mulberry," $F(2, 242) = 1.19$, *ns*, "Mice," $F(2, 242) = .20$, *ns*, or "Softly," $F(14, 1694) = 1.44$, *ns*, with effect sizes less than or equal to 0.01. In addition, no interaction contrasts were significant. Altogether, results implied that Repetitions 2 and 3 for "Mulberry" and "Mice" and Repetitions 2, 3, 5, and 6 for "Softly" could be averaged. In addition, Group was not a factor. Having said that, from Table 2 it can be observed that Repetition 1 was still reasonably consistent with Repetitions 2 and 3, which implies that top-down processing (i.e., GPR 7) was not a major factor in the parsing profiles. More accurately, any effects of top-down processing used in Repetitions 2 and 3 (e.g., scale and tonality extraction) existed in Repetition 1.

Based on these results, the last two repetitions of each stage, for each participant, were averaged to create the individual, average, empirical, boundary profile (hereafter the individual profile). This individual profile was the basis of all subsequent analyses. On a note-by-note basis, for "Mulberry" and "Mice", the individual profile provided three possible values: 0.0 (no boundary), 0.5 (weak boundary), and 1.0 (strong boundary). Because "Softly" averaged Repetitions 2, 3, 5, and 6, the individual boundary profiles could take on five values between 0.0 (no boundary) and 1.0 (strong boundary).

Similarity Analysis

In each stage, the similarity analysis computed the Pearson correlation between the individual profiles of all possible pairs of participants ($123 \times 122/2 = 7503$ correlations) and then computed the mean correlation (with standard deviation). For "Mulberry" (Stage 2, $N = 35$), the mean similarity was $r = .709$ ($SD = .188$, minimum = –.121, median = .740, mode = .938). For "Mice" (Stage 4, $N = 47$), the mean value was $r = .695$ ($SD = .190$, minimum = –.201, median = .730, mode = .762) and for "Softly" (Stages 6 and 8 collapsed, $N = 33$), the mean value was $r = .670$ ($SD = .216$, minimum = –.304, median = .707, mode = 1.000). Maximum values were always $r = 1.00$. As in the consistency analysis, the distributions had some negative skew.

These analyses were supported by detailed cluster analyses for each melody, using the between-groups method with a clustering criterion of $r > .57$ ($r^2 > .33$). This criterion is above the previously cited "intuitive" levels (although, properly, those levels applied only to binary profiles), and it is above the previously cited "significance" levels (comparing each individual profile against $r = 0$). It is comparable to the mean values obtained in the consistency analyses. Each of these cluster analyses found one large group containing more than 90% of the participants, with no other subgroups. That is, participants who were not initially part of the main group did not form unique subgroups; they simply joined the one main group at a lower criterion. In particular, participants from the two boundary efficacy groups did not form into different groups.

Potential differences between the boundary profiles of the two boundary efficacy groups were explicitly tested using a mixed ANOVA, with one between-subjects factor (Group) and one within-subjects factor (Boundary location). It is the interactions that are most interesting. For each of "Mulberry," "Mice," and "Softly," the main effect of Boundary was—not surprisingly—significant, but this is not relevant for the Similarity analysis. For "Mulberry," "Mice," and "Softly," the main effect of Group was never significant. The type of boundary efficacy task did not affect mean performance levels. The Group by Boundary interaction was significant for "Mulberry," $F(34, 4114) = 2.44$, $p < .001$, $\eta^2 = .006$, $\omega^2 = .003$, "Mice," $F(46, 5566) = 2.01$, $p < .001$, $\eta^2 = .005$, $\omega^2 = .003$, and marginally nonsignificant for "Softly," $F(32, 3872) = 1.41$, $p < .063$, $\eta^2 = .004$, $\omega^2 = .001$.

The significance of the interactions involving Group would imply that groups could not be collapsed, whereas the effect sizes for those same interactions would imply that groups could be collapsed. The combination of significance, small *F*s, and trivial effect sizes (i.e., η^2) is due to the enormous power available in a within-subjects design having 4000 degrees of freedom per error term. As a further check, the average boundaries for each group were compared using a Pearson correlation. For "Mulberry," the average boundary profiles of the two groups were essentially the same, with $r = .990$ ($p < .001$). The situation was the similar for "Mice" ($r = .985$, $p < .001$) and "Softly" ($r = .993$, $p < .001$). The cluster analyses, effect sizes, and average correlations implied that the two groups produced the same boundary profiles for each melody.

It must be emphasized that the similarity and consistency analyses do not imply that all participants produced *identical* profiles: These analyses simply validated the use of grouped data for subsequent tests of the model, while implying that individual differences may be considered secondary. Hence, for subsequent analyses, all participants in all groups were

considered as one single group (for "Softly," analyses also collapsed Stages 6 and 8).

Comparisons with *GTTM*

An average boundary profile was created for each stage, for each experiment, by averaging the individual profiles for all participants. This average profile was then compared with the quantified GPRs. The first analysis examined each rule in isolation by using a correlational analysis. The second analysis examined the combination of all rules by using multiple regression (essentially, Intensification: GPR 4). Following the group analyses, the third analysis explored individual differences. For all analyses, attention must be directed to those occasions when a rule fails to predict an empirical boundary (hereafter a *miss*). Note that a single rule may generate several misses. However, given the structure of *GTTM*, every empirical boundary must be associated with at least one low-level GPR. That is, the combination of all rules must not produce any *collective misses*. On the other hand, each rule and the combination of all rules may generate many unfulfilled predictions (hereafter a *false alarm*). These are of little consequence to the validity of the theory, although they affect the correlations between the rules and the empirical data. In this work, the terms miss and false alarm reflect the notion that the *GTTM* is trying to predict the empirical responses. Hence, a miss is an occasion when the theory fails to predict the event and a false alarm is an occasion when the theory predicts an event that did not occur. Note that in the following discussion of the results, references to an event as a miss, false alarm, or correct prediction, or as a strong or weak boundary, are intended only as guides for interpretation and inspection. That is, these labels were not intended to have scientific rigor. The scientific rigor is found in the statistical analyses that treated theoretical predictions and empirical values as continua, in a consistent manner across all melodies and all rules. This analysis did not impose any arbitrary criteria for the classification of events as misses, false alarms, or strong or weak boundaries.

Group Analyses

In the melody "Mulberry" (Stage 2; see Figure 1), participants placed stronger boundaries after Note Events 9, 14, 19, 23, 28, and 35 (i.e., many participants placed boundaries at these points) with weaker boundaries after Note Events 4 and 30 (i.e., fewer participants placed boundaries at these points). Generally, the agreement is high throughout the melody, though there was ambiguity after the strong boundaries and within the last few notes of the melody. Note that the parsing matches the lyrics.

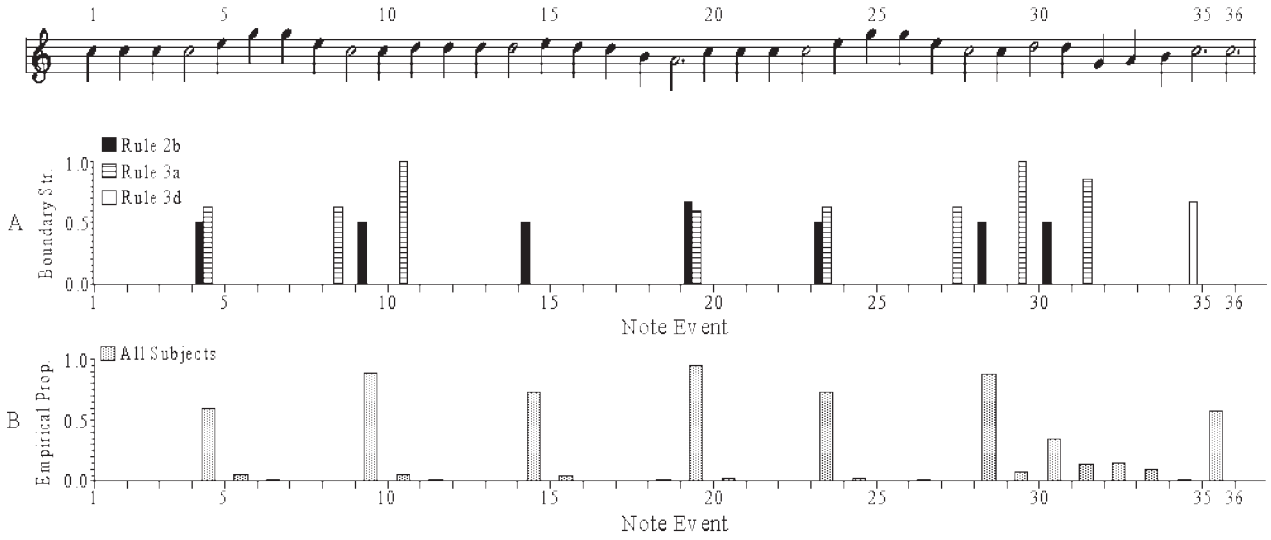

Fig. 1. The melody "The Mulberry Bush," the theoretical boundaries identified by the quantification of GPRs 2b, 3a, and 3d (A), and the empirical boundaries indicated by participants in Experiment 1 (B). The melody was adapted from Bastein (1988), © Neil A. Kjos Music Company, used with permission 2004.

The comparison of Panel A to Panel B in Figure 1 demonstrates the effect of the GPRs. In "Mulberry" (N = 35), Attack-point (GPR 2b) was most successful for predicting boundaries (r = .913, p < .001), capturing essentially all empirical boundaries. Register (GPR 3a) was unsuccessful (r= .157, *ns*). It predicted boundaries on three occasions, but generated false alarms on another four occasions. Length (GPR 3d) was also unsuccessful (r = –.099, *ns*), but Length had only one predicted location. Using multiple regression, the combination of the three rules produced a multiple correlation of R = .913 (p < .001). Given that the value of R is virtually the same as the simple r for Attack-point, it is not surprising that Attack-point was the only significant predictor when a stepwise approach was used in the multiple regression. The conclusion is that Intensification (GPR 4) did not add predictability. Note that the combination of all the GPRs generated a collective miss after Note Event 35. This is a data point that the theory should predict.

In the average profile for "Mice" (Stage 4; see Figure 2), participants placed strong boundaries after Note Events 3, 6, 10 (tied notes represent one event), and 14, with a weaker boundary after Note Event 23. Many locations had what could be called "very minor" boundaries (<20% of the participants indicated boundaries), or more simply "noise." Generally the agreement between participants was high at the beginning of the melody but lower near the end. As with "Mulberry," the parsing matches the lyrics, particularly in the beginning.

For "Mice" (N = 47), as shown in the comparison of Panel A to Panel B in Figure 2, Attack-point (GPR 2b) was the most successful at predicting boundaries (r = .732, p < .001). Register (GPR 3a) was not generally successful (r = .233, ns), but this average hides the fact that it was successful in places (after Note Events 3, 6, 10, and 14), while generating many false alarms (after Note Events 16, 21, 31, and 36). Note that Length (GPR 3d) did not apply within this melody. The combination of the two rules (Intensification: GPR 4) produced a multiple correlation of R = .733 (p < .001). Only Attack-point was a significant predictor when a stepwise approach was used. By inspection, one can see that there might be a collective miss after Note Event 33 or 34 (depending on the arbitrary criterion one might use to designate an empirical boundary).

The average boundary profile for "Softly" (Stages 6 and 8 collapsed) is shown in Figure 3. Generally, participants placed stronger boundaries at locations 9, 16, and 25, with much weaker boundaries at locations 26, 27, and, possibly, 32 and 33. In comparison to the previous melodies, there is more noise throughout the melody (i.e., boundaries are indicated by some participants at each Note Event).

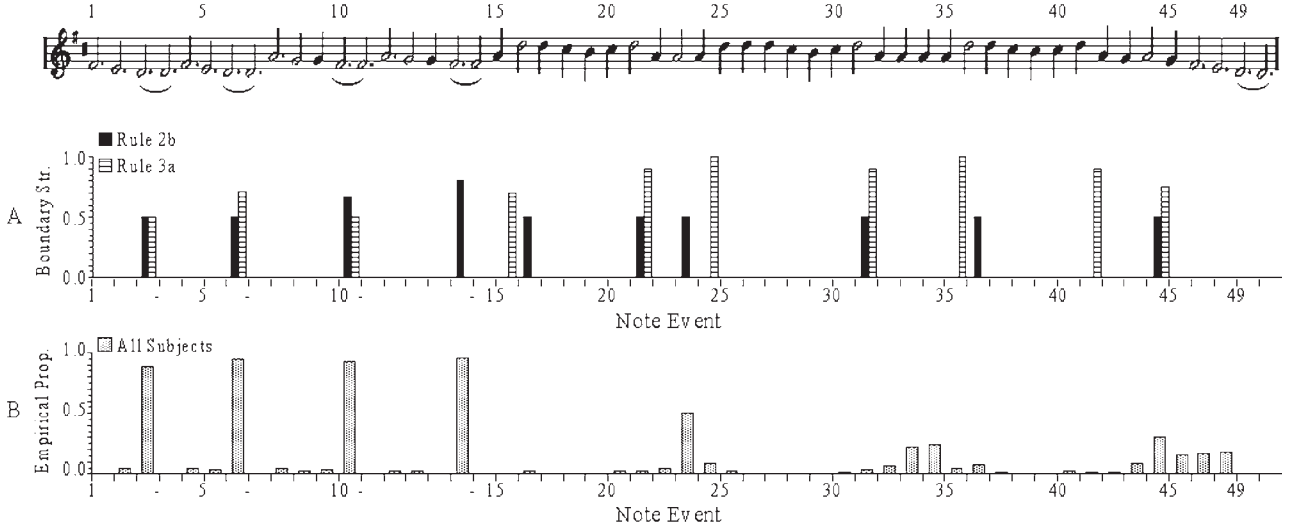

Fig. 2. The melody "Three Blind Mice," the theoretical boundaries identified by the quantification of GPRs 2b and 3a (A), and the empirical boundaries indicated by participants in Experiment 1 (B). The melody was adapted from Bastein (1988), © Neil A. Kjos Music Company, used with permission 2004.

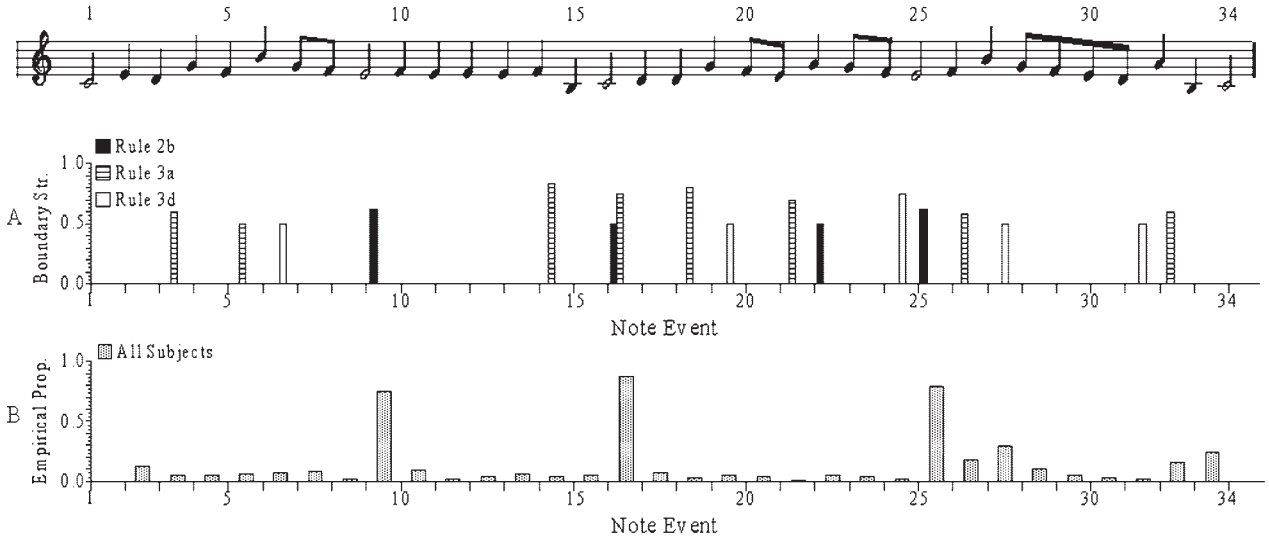

Fig. 3. The extract from the melody "Softly Now the Light of Day," the potential boundaries identified by the quantification of GPRs 2b, 3a, and 3d (A), and the empirical boundaries indicated by participants in Experiment 1 (B). The melody was adapted from Boltz (1989), who obtained it from *The International Library of Music: Album of the World's Best Home Songs* (The Editorial Board of the University Society, 1964, New York: The University Society Inc.).

For "Softly" (N = 33), as can be seen in the comparison of Panel A to Panel B of Figure 3, Attack-point (GPR 2b) was the most successful rule (r = .831, p < .001). The rule predicted three boundaries, with only one false alarm. Register (GPR 3a) was unsuccessful (r = .107, *ns*), generating seven false alarms and one correct prediction. Length (GPR 3d) was not successful (r = –.050, *ns*), generating three false alarms and one prediction of a minor boundary. The combination of three rules (Intensification: GPR 4) produced a multiple correlation of R = .831 (p < .001). Again, only Attack-point was significant as a predictor when using a stepwise approach.

Generally, in all three melodies, the results imply that only Attack-point (GPR 2a) was consistently important. Register (GPR 3a) and Length (GPR 3d) had minimal contributions. The combination of all rules did not improve prediction. In fact, including all two-way interactions between the rules in the multiple regression analyses (using the forced entry of all variables, or the stepwise solution) did not improve prediction beyond Attack-point alone. This implies that, overall, Intensification (GPR 4) was not a useful predictor. However, it was still possible that Register and Length did not add to the equation because their predictions overlapped with those of Attack-point. Table 3 presents the intercorrelations between the rules within each melody (Lerdahl & Jackendoff, 1983, p. 67, call this "confluence"). With the exception of Attack-point and Register in "Mice," the correlations are not significantly different from zero. Hence, it must be concluded that Register and Length were not particularly effective in these melodies.

Individual Analyses

To explore individual differences, the correlation between each rule and the individual boundary profile of each participant was computed. These were labeled *utilities* to distinguish them from previous (average or group) correlations. Each participant produced one correlation per applicable GPR, and therefore, three correlations per melody (only 2 for "Mice" in

TABLE 3
Correlations Between the GPR Within "Mulberry" (Stage 2), "Mice" (Stage 4), and "Softly" (Stages 6 and 8) for Experiment 1

		Correlations Between Measures	
Melody	**GPR**	**Register 3a**	**Length 3d**
Mulberry	Attack-point 2b	.168	–.085
	Register 3a		–.090
Mice	Attack-point 2b	.365**	
Softly	Attack-point 2b	.005	–.137
	Register 3a		–.206

*p < .05. **p < .01.

Stage 4, since Length did not apply). Table 4 provides the average utilities across participants for each rule, for each melody. Table 4 also presents the number of participants who had a correlation above significance (see the previous consistency analysis), as well as the Larzelere and Muliak test that protects Type 1 error rate by using a Bonferroni correction (Howell, 2002). Generally, the conclusions are similar to those of the group analyses, but note that the *maximum* utilities for Register (GPR 3a) and Length (GPR 3d) never reach the level of the *mean* utilities for Attack-point. This analysis also implies slightly greater use for Register than for Length. Furthermore, almost none of the participants had significant utilities for Register (GPR 3a) or Length (GPR 3d).

In summary, all analyses indicated that Attack-point (GPR 2b) was important for parsing, but that Register (GPR 3a) and Length (GPR 3d) were of questionable utility, even though there was considerable opportunity for their use.

Experiment 2

Experiment 2 was a replication of Experiment 1, Group 1 (using the recognition-memory task) using different stimuli with the intent of broadening the base for inference and extending the application to an example from the classical repertoire of music. All elements of the design, including the quantification, were retained in order to maintain comparability across experiments. However, given the consistency within participants shown in the earlier study, only 33 participants were tested. All the analyses of Experiment 1—within-subject consistency, between-subject similarity, comparisons with the GPRs (group and individual)—were conducted but only a subset is presented.

TABLE 4
Statistics Related to the Correlations Between the Individual Boundary Profiles and the GPRs (Utilities) Within "Mulberry" (Stage 2), "Mice" (Stage 4), and "Softly" (Stages 6 and 8) for Experiment 1

Melody	GPR	*M*	*SD*	Min	Max	# Sig[a]	L-M[b]
Mulberry	Attack-point 2b	.768	.146	.333	.992	122	117
	Register 3a	.136	.102	–.116	.444	7	0
	Length 3d	–.083	.035	–.237	.130	0	0
Mice	Attack-point 2b	.610	.109	.090	.795	121	120
	Register 3a	.195	.083	–.085	.483	6	0
Softly	Attack-point 2b	.682	.180	.071	.926	116	103
	Register 3a	.090	.088	–.282	.323	0	0
	Length 3d	–.044	.117	–.278	.288	0	0

[a]# Sig is the number of participants out of 123, who produced a significant correlation ($p < .05$) comparing the individual profile with each GPR.

[b]L-M is the Larzelere and Muliak test that applies a Bonferroni correction to # Sig.

METHOD

Participants

There were 33 participants (18 females) with a mean age of 25.00 (*SD* = 8.27) years (range = 18–48 years) recruited under the same conditions as the previous experiment. These participants had, on average, the equivalent of 6.1 (*SD* = 8.1) years of instruction at 1 hour per week, with RCM or equivalent grades ranging from 0 to 8.

Procedure

The procedure was identical to that of Experiment 1.

Apparatus and Stimuli

The apparatus and setup were the same as those used in Experiment 1: only the stimuli changed. In Stage 2, the melody presented was "Mary Had a Little Lamb" (Figure 4), with notes in the range C_5 to G_5 (where A_4 = 440 Hz), with a tempo of 145 beats per min (quarter note = 414 ms). The computed key strength was r^2 = .34 for the best fitting key (E minor) and a key strength of r^2 = .28 for the second best key (C major), yielding a *q* factor of 0.05 (see Experiment 1). The third best key had a strength of r^2 = .25 (G major). These keys are clustered (circle of fifths), so it could be said to be tonal, although not as strongly as the others used in this work. It had a high familiarity for participants achieving a mean rating (scaled from 0 to 10) of 9.59 (*SD* = 1.09, median, 10.00; mode, 10; range = 5 to 10), and 25 participants could name it. Equipment failure caused the loss of Repetitions 2 and 3 for one participant, so analyses of Stage 2 were based on 32 participants. Hereafter, this melody is referred to as "Mary."

In Stage 4, the melody presented was "Tom, Tom, the Piper's Son" (Figure 5), with notes that ranged from D_4 to D_5. The tempo was set to 135 beats per minute (quarter note = 353 ms). Key strength was r^2 = .79 for the best fitting key (G major) and r^2 = .46 for the second best key (E minor), producing *q* = 0.33. This melody is unambiguous in its tonal center. It had a low familiarity, generating a mean rating of 1.58 (*SD* = 2.18, median = 0.25, mode = 0, range = 0 to 7). No one could name it. Hereafter, it is referred to as "Tom."

In Stages 6 and 8, the melody presented was "Melody in F" (the first theme from Op. 3, No. 1 for piano) by Anton Rubenstein (Figure 6), with notes ranging from $F\sharp_3$ to A_4 (the melody is notated an octave higher in Figure 6), at a tempo of 170 beats per min (quarter note = 353 ms). The melody was transposed to the key of C, for reasons that are not relevant to the current discussion. Key strength was r^2 = .80 for the best fitting key C major) and r^2 = .66 for the second best key (G major), producing *q* = 0.14, implying that it is tonal and unambiguous in key. This melody was unfamiliar, achieving a mean rating of 1.29 (*SD* = 2.29, median = 0.00, mode = 0, range = 0 to 8) in Stage 6. No one identified it. It was repeated in Stage 8 for consistency with Experiment 1. Hereafter, it is referred to as "Melody in F."

RESULTS AND DISCUSSION

Participants produced three boundary profiles in each of Stages 2, 4, 6 and 8 (one boundary profile per repetition of the melody). In each stage, each empirical boundary profile was converted to a binary coding of the presence of boundaries on a note-by-note basis (i.e., a value of 1 indicated a boundary at that note; a value of 0 indicated no boundary at that note).

Consistency Analysis

For the three repetitions for "Mary" (Stage 2; $N = 52$), the overall consistency considering all three repetitions was $r = .592$ ($SE = .152$).[2] For the comparison of Repetitions 2 and 3 alone, the mean consistency was $r = .767$ ($SD = .258$). Both values were different from $r = 0$ ($p < .001$). For Repetitions 2 and 3, 30 of the 32 participants produced correlations that exceeded the critical value of $r = .27$ ($p < .05$); 28 of 32 were significant when using the Larzelere and Muliak test (Howell, 2002) that applies a Bonferroni correction to these values. A within-subjects ANOVA of the three correlations followed by post-hoc contrasts, using a Bonferroni correction (with $p < .025$), demonstrated that the correlation between Repetitions 2 and 3 was higher than the average of the correlations between Repetitions 1 and 2 and Repetitions 1 and 3, $F(1, 31) = 17.44$, $p <. 001$. As in Experiment 1, all analyses comparing correlations used the Fisher r to z transform.

For the melody "Tom" (Stage 4; $N = 29$), the overall consistency was $r = .511$ ($SE = .098$). For Repetitions 2 and 3 alone, the mean consistency was $r = .612$ ($SD = .306$). Both were different from $r = 0$ ($p < .001$). For Repetitions 2 and 3, 26 of the 33 participants produced correlations that exceeded the critical value of $r = .37$ ($p < .05$); 21 of 33 were significant when using the Larzelere and Muliak test. A within-subjects ANOVA followed by post-hoc contrasts, using a Bonferroni correction ($p < .025$), showed no differences between the correlation of Repetitions 2 and 3 and the average of the correlations between Repetitions 1 and 2 and Repetitions 1 and 3, $F(1,32) = 4.53$, $p <. 041$.

For "Melody in F" ($N = 37$; the terminal rest was not included), the overall consistencies were $r = .407$ ($SE = .031$) in Stage 6 and $r = .490$ ($SE = .079$) in Stage 8. In Stage 6, for Repetitions 2 and 3 alone, the mean consistency was $r = .442$ ($SD = .250$); in Stage 8, for Repetitions 5 and 6, it was $r = .581$ ($SD = .260$). All mean values were different from $r = 0$ ($p < .001$). In Stage 6, 21 of the 33 values exceeded the critical value ($r = .33$); when using the Larzelere and Muliak test, the number was 14 of 33. In Stage 8, the corresponding values were 28 and 19 of 33. A within-subjects ANOVA of the 15 correlations (both Stages 6 and 8) using post-hoc contrasts with a Bonferroni correction ($p < .0036$) demonstrated no significant differences between the correlations, although the consistencies involving only Repetitions 2, 3, 5, and 6 were always higher than any involving Repetitions 1 or 4.

Generally, the consistency results were similar to, though lower than, those of Experiment 1. Based on these results, for "Mary," Repetitions 2 and 3

2. Standard errors are reported when discussing a mean of means. Standard deviations are reported when discussing a single mean.

were averaged to create the individual profile for each participant. Although results for "Tom" implied that Repetitions 1, 2, and 3 could be averaged, for compatibility with Experiment 1, only Repetitions 2 and 3 were averaged. Similarly, for "Melody in F," results implied that Repetitions 1 through 6 could be averaged, but for compatibility, only Repetitions 2, 3, 5, and 6 were. These individual boundary profiles were the basis for all subsequent analyses.

Similarity Analysis

When comparing across participants within each stage, for the melody "Mary" (Stage 2; $N = 52$), the mean between-subject similarities were $r = .647$ ($SD = .367$), and for the melody "Tom" (Stage 4; $N = 29$), the mean similarities were $r = .506$ ($SD = .331$). Cluster analyses indicated that, for "Mary" (Stage 2) and for "Tom" (Stage 4), all participants belonged to one large group.

In "Melody in F" (Stages 6 and 8 collapsed, $N = 37$), the mean similarities were $r = .468$ ($SD = .192$). Note that this value is lower than for all other melodies. The cluster analysis, with the same criterion of $r > .57$ ($r^2 > .33$) as used in Experiment 1, resulted in eight groups, but only two had sufficient participants to be useful (hereafter: Group I with $n = 6$ and Group II with $n = 16$). The mean similarities were $r = .707$ ($SD = .124$) in Group I, and $r = .677$ ($SD = .109$) in Group II. Note that within each group, the mean similarities are comparable to other melodies. The boundary profiles for the two groups were analyzed using a mixed ANOVA, with one between-subjects factor (Group) and one within-subjects factor (Boundary). Boundary was significant, but this is not of interest at this time. Group was significant, $F(1, 22) = 93.85$, $p < .001$, $\eta^2 = .339$, $\omega^2 = .334$, implying that Group II placed more boundaries on average than did Group I. The Group by Boundary interaction was significant, $F(36,792) = 7.37$, $p < .001$, $\eta^2 = .274$, $\omega^2 = .236$. The average profiles for Groups I and II were correlated at $r = .550$ ($p < .001$). Using between-subjects *t*-tests, Groups I and II differed on equivalent years of instruction (1.5 vs 8.3 years at 1 hour per week; $t(20) = 2.67$, $p < .016$, maximum intensity of lessons (1.0 vs 2.6 hours per week; $t(20) = 2.23$, $p < .038$, and recency of training (14.0 vs 5.7 years in the past; $t(16) = 3.11$, $p < .007$). Two participants from each group did not train on any instrument. Separate analyses were conducted using all participants and within Groups I and II.

Comparisons With the *GTTM*

The average boundary profile for each stage was created by averaging the individual profiles for all participants in each stage. In addition, for "Melody in F," average boundary profiles were created for Groups I and II separately.

Group Analyses

In the average boundary profile for "Mary" (Stage 2; Figure 4), participants placed strong boundaries after Note Events 7, 10, 13, 26, 34, 37, and 40 with weaker boundaries after Note Event 20 and, perhaps, 47. There was some noise at almost every point (i.e., boundaries indicated by some participants). Note that the parsing matched the lyrics.

For "Mary" (N = 52), Attack-point (GPR 2b) was successful (r = .973, p < .001). Register (GPR 3a) had some success (r = .429, p < .001). From Figure 4 (Panels A and B), one can see that Attack-point and Register applied at many of the same points, but Attack-point did not generate as many false alarms. The combination of both rules (Intensification: GPR 4) produced a multiple correlation of R = .975 (p < .001). Attack-point was the only significant predictor in the equation when a stepwise analysis was used.

In the average boundary profile for "Tom" (Stage 4, Figure 5), participants placed strong boundaries after Note Events 6, 14, 18, and 22 with a weaker boundary after Note Event 2. There were numerous minor boundaries. The empirical boundary profile corresponded to the parsing of lyrics, but this melody was not familiar to participants. As such, the parsing is consistent with the lyrics, but not dependent on knowledge of the lyrics.

In "Tom" (N = 29), Attack-point (GPR 2b) was the most successful rule (r = .914, p < .001), Register (GPR 3a: r = .102, *ns*) and Length (GPR 3d: r = –.077, *ns*) were unsuccessful. As in "Mary," every application of Attack-point was matched to an empirical boundary (compare Panels A and B), but this was not true for either Register or Length. Intensification (GPR 4) produced a multiple correlation of R = .915 (p < .001). Only Attack-point was a significant predictor in the stepwise solution. Note that there is a possibility of a miss after Note Event 2, which is a concern for the theory.

The average boundary profile for "Melody in F" (Stages 6 and 8 collapsed) for all participants is shown in Panel B of Figure 6. Participants placed relatively stronger boundaries after Note Event 10 (effectively, the rest) with somewhat weaker boundaries after Note Events 4, 7, 15, 20, 25, 28, 31, and 36. There were numerous minor boundaries.

For "Melody in F" (N = 37), Rest (GPR 2a) was the most successful (r = .567, p < .001). Attack-point (GPR 2b) was also successful (r = .442, p < .006). Register (GPR 3a: r = –.195, *ns*) and Length (GPR 3d: r = –.144, *ns*) were unsuccessful. For this melody, the combination of the four rules produced a multiple correlation of R = .747 (p < .001). Both Rest and Attack-point were significant predictors in these equations when using a stepwise approach, implying some role for Intensification (GPR 4).

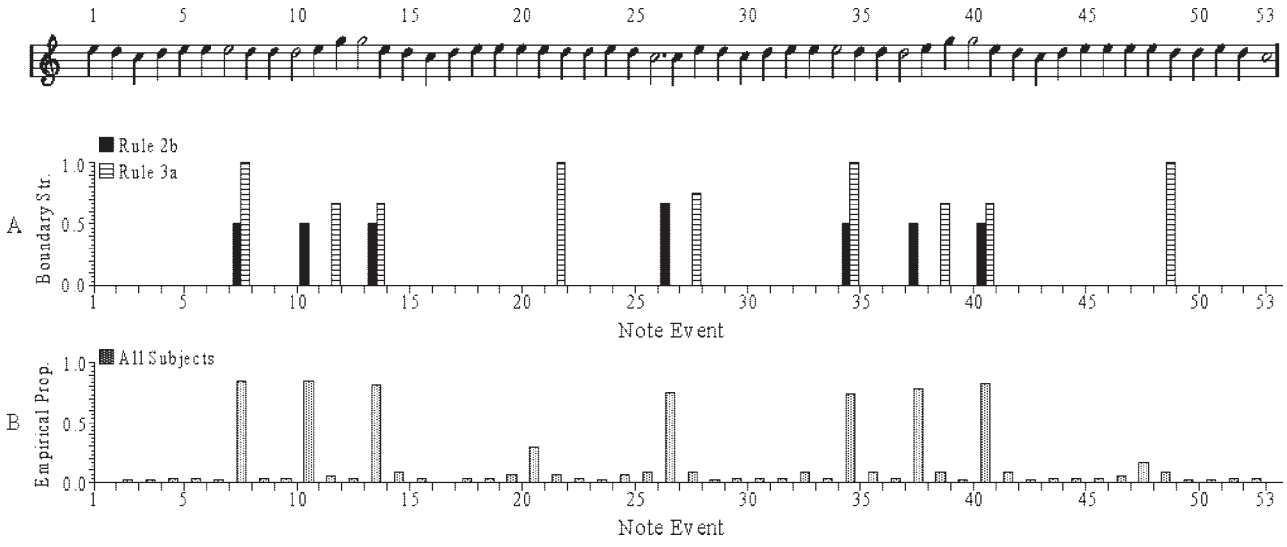

Fig. 4. The melody "Mary Had a Little Lamb," the theoretical boundaries identified by the quantification of GPRs 2b and 3a (A), and the empirical boundaries indicated by participants in Experiment 2 (B). The melody was adapted from Bastein (1988), © Neil A. Kjos Music Company, used with permission 2004.

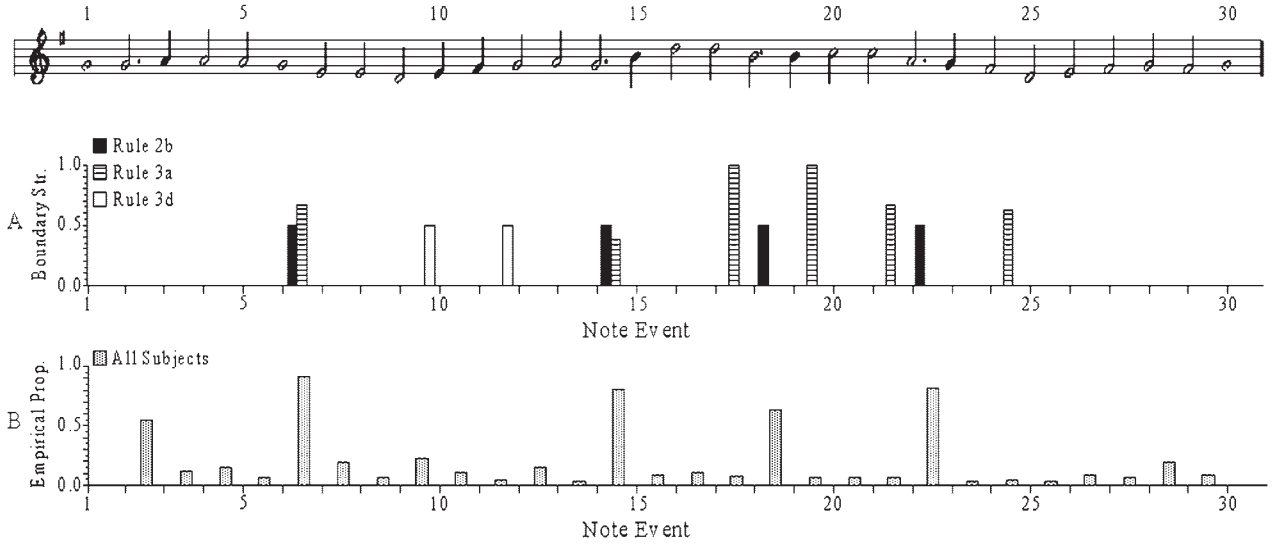

Fig. 5. The melody "Tom, Tom, The Piper's Son," the theoretical boundaries identified by the quantification of GPRs 2b, 3a, and 3d (A), and the empirical boundaries indicated by participants in Experiment 2 (B). The melody was adapted from Bastein (1988), © Neil A. Kjos Music Company, used with permission 2004.

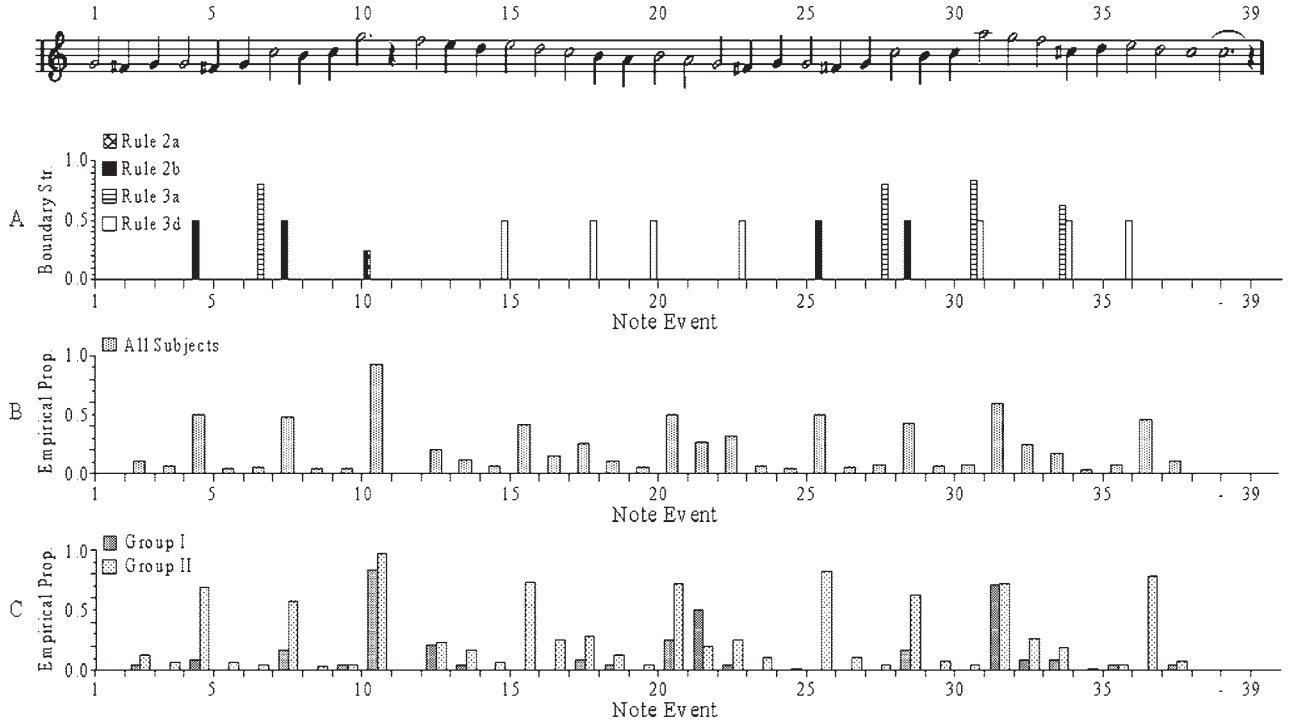

Fig. 6. The melody "Melody in F" the theoretical boundaries identified by the quantification of GPRs 2a, 2b, 3a, and 3d (A), the empirical boundaries indicated by all participants in Experiment 2 (B), and by Groups I and II (C). The melody was adapted from Bastein (1980), © Neil A. Kjos Music Company, used with permission 2004.

Panel C of Figure 6 provides the average boundary profiles for Groups I and II. For Group I, Rest (GPR 2a) was most successful (r = .652, p < .001), while Attack-point (GPR 2b: r = .020, *ns*), Register (GPR 3a: r = –.140, *ns*) and Length (GPR 3d: r = –.147, *ns*) were not. The combination of all four rules produced a multiple correlation of R = .665 (p < .001), but only Rest was a significant predictor when using a stepwise approach. For Group II, Rest (GPR 2a: r = .413, p < .001) and Attack-point (GPR 2b: r = .511, p < .006) were significant, while Register (GPR 3a: r = –.224, *ns*) and Length (GPR 3d: r = –.218, *ns*) were not. The combination of all four rules produced a multiple correlation of R = .692 (p < .001). Both Rest and Attack-point were significant predictors when a stepwise approach is used.

Overall, the analyses of "Melody in F" demonstrated collective misses after Note Events 15, 20, 31, and 36, which are serious concerns for the theory. No empirical boundary should exist without at least one corresponding rule. It is tempting to think that Length (GPR 3d) could apply if one simply shifted the application of Length by one note to the right (i.e., locating the boundary between n_3 and n_4, instead of n_2 and n_3; see Table 1) thereby, aligning four of the seven applications of Length with the missed boundaries. However, though tempting, this would not be Length as defined within *GTTM*. Possible modifications to the theory will be discussed later.

Generally, as in Experiment 1, the results implied that Attack-point (GPR 2b) was consistently the most important rule, although Rest (GPR 2a) was also valid (but there were little data for Rest). Register (GPR 3a) and Length (GPR 3d) had minimal predictability. Even in the multiple regression analysis, Register and Length did not add predictability above that of Attack-point and/or Rest. As in Experiment 1, the inclusion of all two-way interactions between the rules into the multiple regression analyses did not improve prediction or provide any further role for Register or Length. As before, the intercorrelations between the rules, in each melody, were computed (Table 5). Although some correlations are nonzero, they are not high. Therefore, it must be concluded that the lack of effects for Register and Length is not due to overlap with Attack-point or Rest.

Individual Analyses

As in Experiment 1, to examine possible individual differences, the correlation between each rule and the individual boundary profile of each participant was computed (labeled utilities). Table 6 provides the average utilities for each rule for each melody (Groups I and II were not analyzed separately). The results mimic the global analysis. As in Experiment 1, the maximum utilities of Register (GPR 3a) and Length (GPR 3d) generally only approach the mean utility for Attack-point (GPR 2b), implying that

TABLE 5
The Correlations Between The GPR Within the Melodies of "Mary" (Stage 2), "Tom" (Stage 4), and "Melody in F" (Stages 6 & 8) for Experiment 2

		Correlations Between Measures		
Melody	**GPR**	**Attack-point 2b**	**Register 3a**	**Length 3d**
Mary	Attack-point 2b		.383**	–.085
Tom	Attack-point 2b		.144	–.109
	Register 3a			–.132
Melody in F	Rest 2a	–.055	–.054	–.076
	Attack-point 2b		–.114	–.158
	Register 3a			.260

*$p < .05$. **$p < .01$.

TABLE 6
Statistics Concerning the Correlations Between the Individual Boundary Profiles and the GPRs (Utilities) Within "Mary" (Stage 2), "Tom" (Stage 4), and "Melody in F" (Stages 6 and 8) for Experiment 2

Melody	**GPR**	***M***	***SD***	**Min**	**Max**	**# Sig[a]**	**L-M[b]**
Mary	Attack-point 2b	.791	.301	–.333	.993	30	30
	Register 3a	.331	.158	–.165	.508	23	9
Tom	Attack-point 2b	.662	.312	–.229	1.000	29	29
	Register 3a	.079	.148	–.284	.563	1	1
	Length 3d	–.055	.113	–.236	.243	0	0
Melody in F	Rest 2a	.429	.148	.131	.713	23	10
	Attack-point 2b	.276	.242	–.119	.615	16	6
	Register 3a	–.131	.100	–.301	.129	0	0
	Length 3d	.095	.183	–.381	.339	1	0

[a]# Sig is the number of participants out of 33, who produced a significant correlation ($p < .05$) comparing the individual profile with each GPR.
[b]L-M is the Larzelere and Muliak test that applies a Bonferroni correction to # Sig.

these rules are used much less often. Rest (GPR 3a) in "Melody in F" is similar to Attack-point. Individually, most participants had significant utilities for Attack-point but not for Register or Length.

General Discussion

In these experiments, individuals representing a wide range of music backgrounds were able to parse a variety of melodies using a direct online procedure developed for this purpose. Within-subject analyses demonstrated that most individuals were quite consistent across repetitions, becoming more consistent with repetition. Individuals were also more consistent with familiar tunes. Between-subject analyses indicated that all listeners parsed most of the melodies in similar, though not iden-

tical, manners. Training was associated with differences in parsing for only one melody taken from the classical repertoire. These results are similar to those of Clarke and Krumhansl (1990), Deliège (1987, 1989), Deliège and El Ahmadi (1989), Krumhansl (1996), and Peretz (1989), although previous authors did not quantify consistency or similarity (in Experiments 1 and 2 of Deliège, 1989, correlations between participants ranged from $r = .47$ to $r = .92$ depending on the analysis, but details were not provided).

This detailed consistency analysis also supported the notion that the parsing of listeners on the first—or only—pass through a melody would be sufficient to produce a reasonably accurate portrayal of parsing. However, higher consistencies were observed between the second and third repetitions. Experiment 1 also examined the effects of encoding instructions on parsing. The secondary task for Group 1 required participants to memorize the melodies while parsing, while the secondary task for Group 2 did not require memorization. Despite this critical difference, the results from the two groups were essentially identical. The present study supports the assumption that parsing is an automated process that is not easily affected by experimental task. This equivalence is crucial for many studies that have used (or plan to use) the parsing task as a precursor to other experiments (e.g., Clarke & Krumhansl, 1990; Deliège, 1989; Deliège, et al., 1996).

The main purpose of the three experiments was the test of four GPRs of *GTTM* (Lerdahl & Jackendoff, 1983): Rest (GPR 2a), Attack-point (GPR 2b), Register (GPR 3a), and Length (GPR 3d). Of the four, Attack-point was found to have strong, consistent, empirical verification. Rest was also important, but it only applied on one occasion. Predictions based on Attack-point were correlated with empirical boundaries at an $r \geq .71$ (accounting for 50% of the variance) in all melodies except "Melody in F." Correlations were lower for "Melody in F" because Rest also exerted a powerful effect. In all melodies, Register and Length were correlated at a much lower level, with $r \geq .19$ (accounting for less than 4% of the variance) with the exception of Register in "Mary" ($r = .43$). The analysis for individual participants reaffirmed these findings: Attack-point seemed to be used by all listeners to some degree, while Register and Length were not used very much, if at all. No participant used Register or Length as much as Attack-point. In addition, the combination of all rules using multiple regression (Intensification: GPR 4) implied that only Attack-point was useful for predicting boundaries in these melodies (Rest was important in the one melody that used it). Register and Length were not needed.

The strong effect for Attack-point echoes the findings of Deliège (1987). Peretz (1989) did not examine Attack-point, so comparisons can-

not be made directly. However, the lack of strong effects for Register and Length runs somewhat counter to the observations of both Deliège (1987) and Peretz (1989). Deliège (Experiment 1) found that both musicians and nonmusicians parsed in accordance with Register (75% and 48%) and Length (70% and 30%). For Peretz, both musicians and nonmusicians used Register (65% and 54%) and Length (100% and 83%). In these previous studies, all rules seemed to have had some impact. However, when using a design that placed the rules in competition with each other, Deliège (Experiment 2) found a different pattern for Register (48% and 75%) and lower use of Length (33% and 27%).

Difficulties with the interpretation of Deliège (1987) and Peretz (1989) have already been detailed as part of the rationale for the present study, but reiterating those difficulties can help to explain the differences. Deliège (Experiment 1) asked participants to parse complex extracts purportedly selected to contain a single rule. However, it is possible that such extracts contained other rules that were also used for parsing, thereby inflating the attribution of effects. Deliège (Experiment 2) asked participants to parse monophonic sequences designed to contain two competing rules. Since participants had to choose a parsing, the design focused on rules that *could* be used, but not necessarily those that *would* be used in real music. Peretz asked participants to parse monophonic folk melodies, but the stimuli contained numerous opportunities for parsing that were not discussed. Critically, neither Deliège nor Peretz quantified rule strength, which makes it difficult to compare rule use within stimuli, between stimuli, and across experiments. For example, the observed effects of Deliège and Peretz may reflect weak versions of useful rules contrasted with strong versions of less useful rules.

Without quantification (or, at least, precise operational definitions, which would be equivalent), tests of the rules reduce to analysis by intuition (cf. West, Howell, & Cross, 1985). Of course, one may argue that a particular quantification does not capture what was intended by the GPR, but that is an issue of construct validation. Quantification actually encourages debate rather than stifling it. At least, once the rules have been quantified, researchers can clarify the nature of the disagreement.

In this work, the aim of quantification was to remain true to the original definitions of *GTTM* (see Table 1). It now seems reasonable to suggest several improvements in the definitions of these rules. Rests and slurs should not be combined into one rule (Slur/Rest, GPR 2a). Slurs are concerned with the internote interval (the ISI or interstimulus interval) whereas rests are concerned with the absence of a sound for a duration comparable to that of notes, in a position that could be occupied by a note. It would seem more parsimonious to combine the slur with Articulation-change (GPR 3c), which already includes staccato and legato. This per-

spective is consistent with the observation of Deliège (1987) that Rest is an aspect of the score while Slur is an aspect of performance.

As noted previously, to capture the missed boundaries in "Melody in F" (Experiment 3, Stages 6 and 8: Figure 6), the application of Length could be altered to locate a boundary between n_3 and n_4, instead of n_2 and n_3 (see Table 1). Such a change would also improve the alignment of the one application of Length in "Mulberry" (Experiment 1, Figure 1), but it would not help (or hinder) the applications of Length in either "Softly" (Experiment 1, Figure 3) or "Tom" (Experiment 2, Figure 4). However, the same result could be obtained by altering Attack-point (GPR 2b) to capture a boundary after a note that is relatively longer than its predecessors. That is, Attack-point could read (cf. Table 1), "if the interval of time between the attack points of n_3 and n_4 is greater than that between n_1 and n_2 and between n_2 and n_3, then the transition from n_3 to n_4 may be heard as a group boundary." Such an alteration would retain the effectiveness of the current Attack-point in all melodies while capturing most of the missed boundaries in "Melody in F." It would also address the "delayed segmentation" observed by Deliège (1987). This new definition could absorb the function of Length. It is true that dropping the current version of Length would remove the ability to detect a change from a series of long notes to a series of short notes, but this was not the basis of any empirical boundaries in the studied melodies even though there were seven such changes in "Softly" (Figure 3), "Tom" (Figure 5) and "Melody in F" (Figure 6). Parsimony would suggest that functions of Attack-point and Length could be combined into one rule.

The current quantification of Register (GPR 3a) makes it sensitive to small interval changes between adjacent pairs of repeated pitches. As noted previously, this minor problem is due, in part, to the choice of relative scaling (absolute scaling creates different problems). This problem is also due, in part, to the current structure of the GPRs 2 and 3 (see Table 1), which places a boundary between the second and third notes of a four-note span. The limitation to such a short span serves to emphasize relatively small, local effects. Background work for the present experiments using a variety of quantification methods indicated that Register would work better if it compared the current interval to the average of several previous intervals.

Generally, each GPR could be more precisely defined. What is the basis for each rule (e.g., linear or nonlinear scaling)? Should the rules be based on absolute (e.g., the current quantification of Rest) or relative magnitudes (e.g., the current quantification of Register)? It is possible that both absolute and relative codings will be needed for some rules (i.e., two versions of each). These questions also apply to the rules not tested: Dynamic change (GPR 3b) and Articulation change (GPR 3c). Furthermore, in

these definitions, the overlap between rules must be eliminated: The rules should form a set of exclusive disjunctions (cf. the quantification of GPRs 2a and 2b). That is, the rules must have distinct functions even if they generate predications at the same locations. Finally, although internally consistent, it seems that the limitation to exactly four-note spans for all rules is too restrictive, particularly when one considers the range of styles and tempos covered by tonal music. That is, different rules should be allowed to use different spans of notes, as needs or style dictate (cf. Temperley, 2001).

In the refinement of rules, the empirical boundary profiles of the six melodies can be useful. For example, inspection of the figures indicates that, in spite of the previously cited complications, the inclusion of tonality in the low-level rules might be useful, particularly tonality in combination with Attack-point. Most, though not all, empirical boundaries occurred after relatively longer notes that corresponded to the notes of the tonic triad. Krumhansl (1990) has provided substantial evidence indicating that tonally important notes tend to be those that are sounded more often and for longer durations. Deliège (1987, Experiment 2) also noted a tendency to parse melodies on tonally important notes.

As defined, Intensification (GPR 4) seems to refer to the strength of each rule (see Table 1). Because each GPR was quantified on a continuum, that simple definition of Intensification was subsumed within each GPR individually. The definition also implies that Intensification could be considered as the sum of all rules. However, note that there could be additive main effects of GPRs 2 and 3 (with no interaction), and there could be interactions between GPRs 2 and 3 (with or without main effects). It might be more parsimonious to restrict Intensification to the sum of all the interactions of GPRs 2 and 3. Because the interactive interpretation was not specifically advocated by *GTTM*, it was not presented in this analysis (in fact, none of the interactions between the rules improved prediction).

The current work could not be extended to Symmetry (GPR 5) and Parallelism (GPR 6) because these rules are not clearly defined. In fact, *GTTM* admits that the "failure to flesh out a notion of parallelism is a serious gap in our attempt to formulate a fully explicit theory of musical understanding" (Lerdahl & Jackendoff, 1983, p. 53; cf. pp. 51–53). This is unfortunate because it is mainly Symmetry and Parallelism that serve as the link between the low-level rules (i.e., GPRs 2, 3) and the high-level analyses (i.e., Time-Span reduction and Prolongational reduction). In a similar vein, the current work was limited to simple monophonic melodies because the application of the GPR 2 and 3 has yet to be demonstrated in complex homophonic melodies, which constitute a large portion of Western tonal music. It is not sufficient to simply assume, as in *GTTM*, that a single parsing holds for all lines of a homophonic piece. The extension of this type of analysis to homophony will require careful considera-

tion of the applications of different GPRs in different voices, as well as detailed consideration of how such voices can be (if indeed they should be) delineated while listening to music.

It seems prudent to mention that the analysis of the combinations of rules is a nontrivial statistical problem. In a monophonic melody, low-level boundaries might form on the basis of (1) a single rule in isolation, (2) several rules in conjunction, (3) a single rule alone despite the presence of other rules (i.e., several rules could apply, but only one is doing the work), or (4) a single rule reinforced by higher level rules. The situation is even more complex for homophonic music. In addition, the empirical data of an individual (or a group) can indicate parsing at any level of the hierarchy. That is, a listener may create a parsing that corresponds to the lowest level of the hierarchy (GPRs 2 or 3) or to large-scale sections. In principle, one cannot know what level the listener actually used. Instructions to the listener can guide, but cannot determine, the level of parsing. For example, one may ask a listener to parse using only a single low-level rule such as Register. The listener may produce a parsing that includes the unconscious inclusion of other rules or structures such as tonality. Similarly, experimental manipulations may guide, but not determine, the level of parsing, by affecting the depth or focus of parsing (cf. the boundary efficacy tasks of Groups 1 and 2, in Experiment 1).

Given the structure of *GTTM*, every empirical boundary, regardless of the level of parsing used by the listener, must represent *at least one* of GPR 2 or 3. Every higher level boundary must correspond to a low-level boundary. As such, the analysis must watch for a collective miss—an empirical boundary that was not matched to a prediction from *any* GPR. A collective miss implies that the theory is invalid, at least in some part. Some of the melodies used herein demonstrated collective misses. These misses *cannot* be explained by rules not tested: Slur (GPR 2a), Dynamics change (GPR 3b), or Articulation change (GPR 3c). There was simply no information in the stimuli for parsing on the basis of these rules.

On the other hand, every theoretical boundary predicted by each of GPRs 2 and 3 need not appear in the empirical boundary profile. The predictions of GPRs 2 and 3 only indicate *possible* boundaries. In fact, given the theory, most of the predictions of GPRs 2 and 3 will not result in boundaries, particularly at higher levels of the hierarchy. Since the empirical parsing of a listener (or a group) must correspond to some level equal to, or above, the lowest level, a particular prediction may not be manifest in the parsing of the listener. Generally, the higher the level used by the listener, the more false alarms there will be. Even a prediction generated by *all* the low level GPRs together may not manifest at a higher level of the hierarchy. As such, false alarms are not critical to the theory.

Because there is an asymmetry in meaning of misses and false alarms, a new statistical approach was required to address combinations of GPRs 2

and 3. Ordinary least-squares multiple regression cannot handle such an asymmetry. However, this new technique requires an article-length exposition to present the necessary mathematical background and, as such, will form a future manuscript in this series.

Finally, for the purpose of refining the rules, it is noted that participants generally parsed the melodies similarly, regardless of training or internal representation of tonality (assessed in Stages 1, 3, 5, and 7 of each experiment). This generally replicates previous work. Deliège (1987) did find some effects of training, but only for rules that, in her view, pertained to the method of performance (e.g., Slur, Dynamic change, and Articulation change) and not to aspects of score (e.g., Attack-point and Length). Musicians were sensitive to both issues whereas nonmusicians were insensitive to performance issues. In the current work, rules that pertained to performance were controlled and thus could not exert any influence, thereby mitigating against a training effect. In addition, all melodies in the current study were simple and, as such, might fail to provide an opportunity for musically trained individuals to apply their specialized knowledge. The melody having the greater structural complexity, Rubinstein's "Melody in F," in Experiment 2, did lead to effects of training on parsing. The role of training is ambiguous in *GTTM*. The GPRs are based on Gestalt concepts of proximity and similarity, and as such, are thought to tap natural, "idiom independent" processes of auditory pattern perception (Lerdahl & Jackendoff, 1983, p. 36). However, *GTTM* also states that the theory applies only to the "experienced listener" (Lerdahl & Jackendoff, 1983, p. 4), which presupposes that the perceptions of the inexperienced listener may be different. At yet another point, the theory states that "the listener needs to know *relatively little* about a musical idiom to assign grouping structure" (Lerdahl & Jackendoff, 1983, p. 36; italic ours) without actually defining "relatively little." Given the inconsistency, one might surmise that some aspects of the theory are insensitive to training (i.e., certain low-level rules), while others are sensitive to training (i.e., certain low-level rules and the higher level rules).

In general, these two experiments found that the GPRs of *GTTM* had some predictive validity. The results suggest that Attack-point (GPR 2b) be retained, perhaps in a slightly altered form, while other rules be used to capture those few boundaries that the new Attack-point missed. In future work, in addition to refining the rules addressed in the present study, the remaining rules need to be quantified and tested, and new analyses need to be developed that can accommodate combinations of the rules, including the asymmetry between misses and false alarms.[3]

3. This article is based on Chapters 1–3 and 5 of the doctoral thesis of Bradley W. Frankland, entitled Empirical Tests of Lerdahl and Jackendoff's (1983) Low-Level Group Preference Rules for the Parsing of Melody, Dalhousie University, August 1998. Aspects of these results have been presented at the annual meeting of the Canadian Society for Brain,

References

Acker, B., & Pastore, R. (August 1996). Melody perception in homophonic and polyphonic contexts. In *4th International Conference on Music Perception and Cognition: Proceedings* (pp. 453–458). Montreal: ICMPC.

Allen, L. G. (1979). The perception of time. *Perception & Psychophysics, 26,* 340–354.

Bastein, J. (Arranger). (1980). *Classic themes by the masters.* San Diego: Neil A. Kjos Music Company.

Bastein, J. (Arranger). (1988). *Nursery melodies at the piano*. San Diego: Neil A. Kjos Music Company.

Berent, I., & Perfetti, C. A. (1993). An on-line method in studying music parsing. *Cognition, 46,* 203–222.

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

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

Bozzi, P., Caramelli, N., & Zecchinelli, L. (1994, July). Figure-Ground: An experiment on perceptual principles of music organization in simultaneous melodies. In *3rd International Conference on Music Perception and Cognition: Proceedings* (pp. 233–236). Liège, Belgium: ICMPC.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics.* Toronto: Harcourt, Brace, Jovanovich.

Clarke, E. F., & Krumhansl, C. L. (1990). Perceiving musical time. *Music Perception, 7,* 213–251.

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

Cohen, A. J. (2000). Development of tonality induction: Plasticity, exposure and training. *Music Perception, 17,* 437–459.

Cohen, A. J., Trehub, S. E., & Thorpe, L. A. (1989). Effects of uncertainty on melodic information processing. *Perception & Psychophysics, 46,* 18–28.

Cook, N. (1989). Music theory and "good comparison": A Viennese perspective. *Journal of Music Theory, 33,* 117–141.

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

Deliège, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's Grouping Preference Rules. *Music Perception, 4,* 325–360.

Deliège, I. (1989). A perceptual approach to contemporary musical forms (D. Dusinberre, Trans.). In S. McAdams & I. Deliège (Eds.), *Contemporary Music Review: Music and the Cognitive Sciences: Volume 4, Proceedings from the 'Symposium on Music and the Cognitive Sciences,' 14–18 March 1988* (pp. 213–230). New York: Harwood Academic Publishers.

Deliège, I., & El Ahmadi, A. (1989). Mechanisms of cue extraction in musical groupings: A study of perception of Sequenza VI for viola solo by L. Berio. *Psychology of Music, 18,* 18–44.

Deliège, I. Mélen, M., Stammers, D., & Cross (1996). Musical schemata in real-time litening to a piece of music. *Music Perception, 14,* 117–160

Dowling, W. J. (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception & Psychophysics, 14,* 37–40.

Behaviour and Cognitive Sciences, June 1998 (Ottawa, Ont., Canada) and at the Fourth International Conference on Music Perception and Cognition, August 1996 (Montreal, Que., Canada).

Grants from the Natural Sciences and Engineering Research Council (NSERC) to A. J. Cohen and an NSERC postdoctoral fellowship to B. W. Frankland, held while the manuscript was revised, are gratefully acknowledged. Appreciation is expressed to Stephen McAdams (Action Editor), Carolyn Drake, and two anonymous reviewers for their careful critiques and advice.

Frankland, B. W., & Cohen, A. J. (1990). Expectancy profiles generated by major scales: Group differences in ratings and reaction times. *Psychomusicology, 9,* 173–192.

Frankland, B. W., & Cohen, A. J. (1996). Using the Krumhansl and Schmuckler key-finding algorithm to quantify the effects of tonality in the interpolated-tone pitch-comparison task. *Music Perception, 14,* 57–83.

Gregory, A. H. (1978). Perception of clicks in music. *Perception & Psychophysics, 2,* 171–174.

Handel, S. (1993). *Listening: An introduction to the perception of auditory events.* Cambridge, MA: MIT Press.

Howell, D. C. (2002). *Statistical methods in psychology.* Pacific Grove, CA : Duxbury Press.

Jackendoff, R. (1992). Musical processing and musical affect. In M. R. Jones & S. Holleran (Eds.), *Cognitive bases of musical communication* (pp. 51–68). Washington, DC: American Psychological Association.

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

Jusczyk, 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,* 627–640.

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11,* 346–374.

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

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension and musical ideas. *Music Perception, 13,* 401–432.

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,* 579–594.

Lerdahl, F. (1988a). Tonal pitch space. *Music Perception, 5,* 315–350.

Lerdahl, F. (1988b). Cognitive constraints on compositional system. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 231–259). Oxford: Clarendon Press.

Lerdahl, F. (1992). Pitch space journeys in two Chopin preludes. In M. R. Jones & S. Holleran (Eds.), *Cognitive bases of musical communication* (pp. 171–196). Washington, DC: American Psychological Association.

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

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* Cambridge, MA: MIT Press.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality & Social Psychology, 28,* 28–38.

Newtson, D. (1976). Foundations of attribution: The perception of ongoing behavior. In J. Harvey, W. Ickes, & R. Kidd (Eds.), *New directions in attribution research* (Vol. 1, pp. 223–247). Hillsdale, NJ: Erlbaum.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality & Social Psychology, 12,* 847–862.

Palmer, C., & Krumhansl, C. L. (1987). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception & Performance, 13,* 116–126.

Peretz, I. (1989). Determinants of clustering music: An appraisal of task factors. *International Journal of Psychology, 24,* 157–178.

Peretz, I., & Babai, M. (1992). The rule of contour and intervals in the recognition of melody parts: Evidence from cerebral asymmetries in musicians. *Neuropsychologia, 30,* 277–292.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 59,* 305–333.

Sloboda, J. A., & Gregory, A. H. (1980). The psychological reality of musical segments. *Canadian Journal of Psychology, 34,* 274–280.

SPSS users' guide. (1988). Chicago: SPSS, Inc.

Stoffer, T. (1985). Representation of phrase structure in the perception of music. *Music Perception, 3,* 191–220.

Tan, N., Aiello, R., & Bever, T. G. (1981). Harmonic structure as a determinant of melodic organization. *Memory and Cognition, 9,* 533–539.

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

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

West, R., Howell, P., & Cross, I. (1985). Modelling perceived musical structure. In P. Howell, I. Cross, & R. West (Eds.), *Musical structure and cognition* (pp. 21–52). New York: Academic Press.