Statistics, Structure, and Style in Music

Carol Lynne Krumhansl
Cornell University

This essay begins by reviewing issues in psychological measurement that motivated some of the research summarized in *Cognitive Foundations of Musical Pitch* (Krumhansl, 1990). These were challenges to geometrical models of similarity, asymmetrical measures of similarity, and contextual effects. It then considers the impact that statistical learning has had on research and theory about music cognition, suggesting this emphasis may have underestimated the importance of other psychological processes contributing to the experience of music. Finally, it discusses three problems with traditional analyses using unidimensional, sequential statistics. The first is that music is hierarchical, with important relationships between non-adjacent events. The second is that the dimensions of music, specifically, pitch and time, interact. The third is the assumption that probabilities remain constant throughout a composition. Rather, music contains artful deviations from normative probabilities contributing to the experience of tension and release.

Received: December 30, 2014, accepted April 1, 2015.

Key words: psychological measurement, tonality, statistical learning, corpus analysis, tonal and event hierarchies

The problem of measurement is fundamental to psychological science. Whereas the physical sciences have developed precise observer-free measures of physical entities, psychological measurement is comparatively informal and uncertain (Luce & Krumhansl, 1988). At a basic level, the difficulty stems from the problem of quantifying phenomena that are essentially subjective, that is, internal to an individual. Sensations can generally be described in terms of relative magnitudes (unidimensional scaling) but, despite the rigorous methods of psychophysics, external behavioral measures vary, across individuals and within individuals over time. The difficulties compound when attempting to describe subjective experiences that are not directly tied to concrete physical objects, such as words and concepts. In some cases, the psychological representation can be investigated by eliciting similarity judgments, yielding a map of the relative distances between entities in a low-dimensional space (multidimensional scaling).

Two important challenges to psychological measurement came in the late 1970s, and these strongly influenced the research reported in *Cognitive Foundations of Musical Pitch* (Krumhansl, 1990, henceforth CFMP). Tversky (1977) questioned the basic assumptions underlying spatial representations such as those used in multidimensional scaling by demonstrating violations of the metric axioms (minimality, symmetry, and the triangle inequality) in similarity judgments. He proposed an alternative model, the feature-matching model, based on the idea that a stimulus, particularly a conceptual one, has multiple features and the number of features known about an object will vary, affecting similarity judgments. Moreover, the context may shift attention from one set of features to another.

Various proposals (reviewed in CFMP) showed that some of these difficulties, particularly the violations of metric axioms, could be addressed by generalizations of the basic multidimensional scaling model. Krumhansl (1978) proposed the distance-density model that posited that judged similarity is a function, not only of the distance between the objects in a multidimensional space, but also the density of other objects around the two objects being compared. The model could account for quite a wide range of similarity data, but two theoretical points made by Tversky (1977) were especially cogent for the music experiments: the existence of asymmetrical similarity judgments (or memory confusions, or other kinds of data), and the potency of contextual influences.

The second important challenge to traditional ideas of measurement in psychological science came from Rosch (1975, 1978; Rosch & Mervis, 1975). She noted that within perceptual and conceptual categories, certain objects, sometimes referred to as "cognitive reference points" or "prototypes," have special psychological status; they are most typical, normative, or the best exemplars of the category. They are the reference points to which other objects are compared. The empirical support for cognitive reference points (or prototypes) came from two main lines of research. One demonstrated that judges reliably

Music Perception, volume 33, issue 1, pp. 20–31, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2015.33.1.20

order objects within categories in terms of prototypicality, establishing a hierarchy of category members. The hierarchy was also reflected in other measures; more prototypical members tend to be better remembered, have shorter names, and be learned earlier in development. The other line of studies demonstrated asymmetrical patterns in behavioral measures. For example, a less prototypical member is judged as more similar to a more prototypical member, than the opposite. Two theoretical points informed the early music studies: again, asymmetrical measures and hierarchical ordering within categories.

The first probe-tone study (Krumhansl & Shepard, 1979) had the relatively modest goal of determining the hierarchy of tones in a tonal context, analogous to the prototypicality of concepts within categories. In this case, incomplete (the final tonic was missing) ascending and descending major scales were used as the tonal context. The context was followed by each tone of the chromatic scale in turn, the "probe tone." Each probe tone was judged as to how well it completed the sequence. The notion of a tonal hierarchy was suggested by music theory, where it is variously referred to as a hierarchy of stability, structural significance, priority, resolution (version tension), and rest (versus activity). The second probe-tone study (Krumhansl & Kessler, 1982) expanded on this by using a variety of chord contexts and major and minor keys, in addition to scales. The participants judged how well the probe tone "fit with" the context. Another experiment in that paper also probed chord sequences on a step-by-step basis to study how the sense of tonality develops and changes over time.

Owing to practical limits on synthesizing sounds at the time, the basic probe-tone studies used short contexts, scales, chords, or chord cadences to establish the keys. This had the advantage, though, of demonstrating that the different contexts invoked the same abstract, schematic knowledge of musical structure; the different contexts resulted in highly correlated results despite the fact that the specific tones present in the contexts varied. Claims have been made that the probe-tone data can be accounted for by lower-level perceptual processes (for example, Leman, 2000; Large, 2011). However, these models have used a composite of the different contexts from Krumhansl and Kessler (1982) as input; that is, all the tones of the separate contexts were summed together before being input to the model. Had the contexts been entered separately, these lower-level models would not give the high correlations found in the probe-tone data for the different contexts (see Krumhansl & Cuddy, 2010, for further discussion of this point).

Another suggestion that the tonal hierarchy embodied abstract, schematic knowledge came from mathematically deriving a map of musical keys from the probe-tone data. Keys that are theoretically closely related had similar hierarchies. When they were correlated with one another, and the correlations were submitted to multidimensional scaling, the result was a toroidal representation that showed the circle of fifths, and relative and parallel major-minor key relationships. Locally, the scaling result contained Schoenberg's (1906/1978) map of keys. Thus, the data gave rise to something clearly on a theoretical plane.

These results focused attention on the relationships between different levels, in this case between tones and keys. Would similar effects be found for chords? An experiment reported in CFMP found a hierarchy of chords in tonal contexts. The same mathematical steps, of correlating the hierarchies and scaling the correlations, yielded a toroidal map that was virtually identical to that derived from the tonal hierarchy. This was true despite the completely independent judgments of tones and chords and important differences in the hierarchies obtained (for example, the different relative positions of the third and fourth scale tones and the chords built on the third and fourth scale tones). Again, this suggested that these studies were tapping into something cognitive, rather than perceptual, in nature. The hierarchical differentiation of both tones and chords reflects a general tendency to encode and remember elements in term of their relations to a few stable, psychologically central reference points.

Two other measurement issues in the work of Tversky and Rosch were also involved in the music studies. One was the existence of asymmetrical measures and the other was the potency of context effects. Experiments on harmony, many of them done in collaboration with Bharucha (Bharucha & Krumhansl, 1983; Krumhansl, Bharucha, & Castellano, 1982; Krumhansl, Bharucha, & Kessler, 1982), focused on these issues and found similar effects to those that had been found for tones (Krumhansl, 1979). In rating studies of pairs of tones and pairs of chords, higher ratings of similarity were found when the second element was the more stable in the key context than the reverse. The magnitude of the asymmetries could be modeled well by the relative differences in the probe-tone and probe-chord ratings. Similarly in memory studies, less stable and more stable elements were more often confused with one another when the less stable element was heard first; the less stable element was more difficult to encode and remember and was more easily assimilated to the more stable element than the reverse. These asymmetries again fit with music-theoretical descriptions contrasting tension and resolution, activity and rest, and instability and stability.

The contextual effects were also striking, regular, and consistent with music theory. When presented in a tonal context, the stable tones (with higher probe-tone ratings) were perceived as more related to one another than those that are less stable. This means that, when depicted in a spatial representation (ignoring the asymmetries), intervals of the same size vary in their proximity. This argues against the use of spatially regular models of musical relationships in which all intervals of a fixed size are shown as equally distant. The same kind of contextual effects were found for chords, perhaps most clearly in a study by Krumhansl, Bharucha, and Castellano (1982), which varied the key context around the circle of fifths. The shift of context resulted in the predicted systematic deformations in the spatial representation, with the degree to which chords were drawn closer together a function of the distance between their key and the key of the context. These contextual effects were reviewed and summarized theoretically in Bharucha and Krumhansl (1983).

Thus, it is with some justification that Shaffer (2007) in his review of CFMP states, "If the experiments had not confirmed music theory, it is the measurement method rather than the theory that would most probably have been called into question. This takes us to what is likely to be the real point of the studies, that the music theory is being used to test the adequacy of the measurement methods. These are studies of measurement. In the event the methods performed very well, thanks to the virtuosity of the experimenter. *Cognitive foundations* makes exciting reading for anyone interested in measurement technique." (p. 179). Shaffer may, however, have underestimated the significance of the convergence between the empirical data and theoretical descriptions of musical structure. The experiments demonstrated that music theory has considerable psychological validity, at least for the basic structural features investigated empirically. Reflecting on this situation, Hyer (2003) included in his *Grove dictionary* entry on tonality, "A further complication (and recurrent tension) has to do with whether the term refers to the objective properties of the music – its fixed, internal structure – or the cognitive experience of listeners . . . "

Nonetheless, various issues arise concerning how this convergence can be understood. After all, the two disciplines of music theory and music cognition have quite distinct objectives, employ different methods of analysis, and consider in many cases different kinds of musical materials. Music theory is directed at describing circumscribed style systems, with an eye toward analysis and composition, often working from the page and considering extended musical examples. It tends not to address more general issues about why music tends to be structured in certain ways, or how it is processed in real-time. By contrast, music cognition is concerned with uncovering general principles of psychological organization, sometimes tested using what are hoped-to-be representative musical materials sometimes considerably reduced in complexity.

Shaffer (2007) raised a related question, "But in what sense can these scaling exercises explain music theory? A description of mental representation based on data remains a summary of the data, which can be valuable. It becomes an explanatory model only when it also describes the processes that construct the representation and transform it in its use. What Krumhansl has to offer towards this arises from observing the similarity between the rating profiles and the distributions of note usage analysed in samples of music from major composers. She is led to suggest that people learn the statistical properties of tonal music and form tonal expectancies from these." (p. 180). As detailed in CFMP, not only were the perceptual judgments correlated with the statistical frequency of tones, they were correlated with transitions between tones, chord distributions, and transitions between chords.

A study with North Indian *ragas* (Castellano, Bharucha, & Krumhansl, 1984) had earlier obtained a finding related to the convergence between probe-tone ratings and the distribution of tones in Western music. Western and Indian listeners produced strikingly similar probe-tone judgments in response to *raga* contexts, consistent with theoretical predictions for the Indian music. In order to explain the agreement between the two groups of listeners, the contexts were examined. It was found that the theoretically important tones were sounded for longer durations than other tones. Apparently, the unfamiliar listeners relied on this information about the distributions of tones in large part when making their judgments (see, however, Large, Kim, Flaig, Bharucha, & Krumhansl, in press, for additional factors correlating with the judgments). Oram and Cuddy (1995) and Cuddy (1997) conducted experiments with novel tone distributions, and found probe tone ratings were systematically related to the frequency of occurrence of the tones in the sequence. Although prior musical knowledge also played a role in these experiments, the results generally pointed to a flexible system for processing novel pitch distributions.

These and many other related studies on the tonal hierarchy are summarized by Krumhansl and Cuddy (2010), where we concluded: 1) tonal hierarchies have psychological reality – that is, they are represented cognitively and play a central role in how musical sequences are perceived, organized, and remembered and in how expectations are formed, 2) they are musical facts - manifest in the way music is written and how its structure is codified in music theory, and 3) that statistically frequent patterns in the music should, in most cases, be reliable guides to the listener for abstracting the tonal hierarchy.

Initially, the fit between the probe-tone judgments and the statistical distribution of tones in tonal-harmonic music (first reported in Krumhansl, 1985) provided a source of external validation for the experimental results. However, at the same time it seemed odd that the structures identified by music theorists, and shown in the experiments to have psychological reality, were reflected so directly in the low-level statistics. What features of the music were being summarized in these statistical analyses, and how were they learned and mentally represented by the listeners? And, how did this sit with the idea that the rating and scaling data were tapping into abstract, schematic knowledge presumably acquired through long-term exposure to music?

Some assurance that this was to be expected came from Meyer (1956, 1957, 1967), who was strongly influenced by information theory in the 1950s and gave a great deal of thought to the relationship between music and probabilities. "That musical styles are internalized probability systems is demonstrated by the rules of musical grammar and syntax found in textbooks on harmony, counterpoint, and theory in general. The rules given in such books are almost invariably stated in terms of probability. For example, we are told that in the tonal harmonic system of Western music the tonic chord is most often followed by the dominant, frequently by the subdominant, sometimes by the mediant, and so forth." (Meyer, 1967, p. 8). So, perhaps it was not entirely surprising that perceptual judgments corresponded to low-level statistics.

Some support was also found in various traditions in psychology. Psychological research on learning had shown a high degree of sensitivity to frequency information, that is, how often various stimuli are presented and how often these are paired with favorable outcomes. This sensitivity enables the organism, human or animal, to classify stimuli and to respond appropriately. Estes' (1950) influential learning theory gave a statistical interpretation to environmental events, proposing theoretical laws that state probability relations between environmental conditions and behaviors. Gibson (1966, 1979) had emphasized in his ecological theory that the important information for perception can be found in the stimulus itself. The organism is so attuned to these stimulus properties that it is unnecessary to hypothesize additional, mentalistic constructs. Brunswik (1956) had stressed that it is essential to understand the statistical properties of stimuli as they occur singly and in combination in the environment.

The correspondences between perceptual judgments and statistical frequencies of tones, chords, and transitions between them are important because they suggest a mechanism for how knowledge of tonality is acquired. They also fit with the large literature on "statistical learning" in other domains, such as in language and visual perception. However, it should be remembered that many other psychological factors, such as low-level sensory processing, short-term memory limits, the vast capacity of long-term memory, dynamics of attending, associative memory, motor control and planning, and sensory-motor integration, are relevant to music cognition. At a general level, it would seem that the important questions for music cognition are: 1) the nature of the musical information that is important for the listener or performer, 2) how it becomes organized cognitively into a coherent system, and 3) how this organizational framework serves to enhance the appreciation and comprehension of the musical experience.

Principles of perceptual organization, and related Gestalt concepts, would seem especially relevant to processing musical structure. According to Meyer (1967), "The laws of mental behavior, formulated by *Gestalt* psychology, are important shaping forces in composition – and they are not basically statistical." (p. 260). The strength of context in determining the role of constituent elements in a larger form is described in Gestalt theory by " . . . the fundamental 'formula' of Gestalt theory: There are wholes, the behavior of which is not determined by their individual elements, but where the part-processes are themselves determined by the intrinsic nature of the whole." (Wertheimer, 1938, p. 2). Interestingly, some of the earliest insights in Gestalt psychology came from music. Wertheimer (1938) gave the example, "What takes place in each single part already depends upon what the whole is. The flesh and blood of a tone depends from the start on its role in the melody." (p. 5). This tradition influenced the emphasis on contextual factors in CFMP and related experiments cited earlier. Perceptual organization is also fundamental to the research on auditory scene analysis (Bregman, 1990) and is at the core of the implication-realization model of melodic processes (Narmour, 1990).

I would like to raise the possibility that a disproportionate emphasis might currently be given to musical statistics and our sensitivity to them. As of late 2014, a Google Scholar search for the word "music" and exact match to the phrase "statistical learning" turned up more than 7,500 hits. Glancing through the first thousand or so suggested many of the hits were not spurious. The papers come from many different disciplines, including cognitive and developmental psychology, computer science, computational modeling, neuroscience, music information retrieval, music theory, and ethnomusicology. What I would like to do in the remainder of this essay is to raise some specific issues concerning the use of statistical analysis of music, particularly one-dimensional (considering only one dimension of music at a time), sequential statistics. In this, I will draw heavily from Meyer's insights about the limitations of statistics to capture essential properties of musical styles and give insights into human musical capacities.

Meyer (1957) cautions against relying too heavily on statistical summaries. "The mere collection and counting of phenomena do not lead to significant concepts. Behind any statistical investigation must be hypotheses that determine which facts shall be collected and counted." (p. 421). And more recently, "Particularly when computerization forces us to define traits with precision statistics can help us to refine our discriminations among classes and styles. But in the absence of hypotheses about syntactic rules and strategic goals, raw statistical data may lead to meaningless or plain foolish results." (Meyer, 1989, p. 58). He gives the examples, "... the number of horns or the presence of tubas in a symphony orchestra is probably a reasonably reliable marker for distinguishing German symphonies of the late Romantic period from Classic symphonies. But per se (as brute facts) they are trivial traits. They tell us little about the differences between the way each style works ... It may be possible to distinguish Wagner's opera from Verdi's in terms of the relative frequency of deceptive cadences. But unless one can explain how the frequency of deceptive cadences is related to other features of Wagner's style the trait will be no more than a marker." (p. 62).

It is interesting to consider Meyer's cautions from the perspective of the present state of music informatics and "big data." We currently have vastly expanded resources for encoding and analyzing music data compared to any time previously. For example, Zivic, Shifres, and Cecchi (2013) undertook a statistical analysis of an enormous corpus of music from 1730 to 1930. It consists of an historical record of melodic interval counts; they focused specifically on pairs of successive intervals. Their analysis uncovered a set of factors that sharply distinguished between Baroque, Classical, Romantic, and Post-Romantic styles. The factor that distinguished Baroque music from other styles was found to be three adjacent notes along the diatonic scale. Classical music was distinguished from other styles by the number of times a note is repeated three times in succession, that is, note-note-note. The Romantic period showed more intervallic diversity and fewer adjacent scale steps. The post-Romantic period showed an even wider variety of intervals still, also larger in size. Uncovering these defining characteristics is interesting, but one wants to ask why these particular features, what other features are related to these, how they compare with observations from other analytic approaches, and most importantly, whether and why they might be important for music cognition.

Before turning to some specific problems with unidimensional, sequential statistics, it may be well to consider briefly Meyer's theory of how probabilities relate to emotion and meaning in music. According to him, "Out of such internalized probability systems arise the expectations – the tendencies upon which musical meaning is built. But probability is not the same as expectation. Or, to put the matter in another way, we must distinguish between active and latent expectation – between the fact of probability and the awareness that an individual has of alternative probabilities." (Meyer, 1957, p. 414). Probabilities are facts about music; they may be internalized as latent, passive knowledge. Expectations, on the other hand, are active, creating tension when considering alternative future events. "Thus in a very general way expectation is always ahead of the music, creating a background of diffuse tension against which particular delays articulate the affective curve and create meaning." (Meyer, 1956, p. 59). Tension is created when an expected event is delayed or the context is ambiguous. Thus, according to Meyer, when there is no delay or ambiguity and an improbable event simply occurs, there is no tension, no meaning, and no emotion, except perhaps surprise or laughter.

A first problem with statistical analyses of music is that they are generally done on sequential events, as in Markov models. But music is hierarchical in the sense that significant relationships exist between non-adjacent events. Meyer (1989) addressed this problem as follows, "And often the most important relationships are to be found in levels other than that of the "facts" notated in the score. This presents the formidable methodological problem of developing objective and rigorous rules for defining and differentiating patterns above the level of the foreground – the note-to-note level. What are needed are rules for distinguishing structural tones, harmonies, accents, and so on, from ornamental tones on all hierarchical levels." (p. 58). He suggests that perhaps some of the problems may be ameliorated with large enough samples, where the error will be absorbed into the mass of data.

A similar caution about statistical treatments of music comes much earlier from Gestalt psychology, "A melody is a whole, organized in time . . . The earlier notes of the melody have an effect upon the later ones, because they have started a process which demands a definite continuation. A melody . . . [is] . . . not analogous to beads on a string, but. . . . [is] . . . a continuous process . . . These events have their own shape which demands a proper continuation. Without knowing something about these shapes, it is difficult to judge which events are to be counted." (Koffka, 1935, p. 437).

It might be argued that the empirical support found for Narmour's (1990) implication-realization (I-R) model suggests it is possible to develop a theory that has strong predictive value on a note-to-note level. Most of the experiments testing the theory have indeed focused on successive tones (e.g., Cuddy & Lunney, 1995; Eerola, Järvinen, Louhivuori, & Toiviainen, 2001; Krumhansl, 1995; Krumhansl, Louhivuori, Toiviainen, Järvinen, & Eerola, 1999; Krumhansl et al., 2000; Rohrmeier & Cross, 2013; Thompson, Cuddy, Plaus, 1997). However, Narmour's theory also considers implications between non-adjacent tones. His analyses contain numerous examples of implicative relationships over longer spans of time. One stipulation of the model is that when some characteristics of a pattern are repeated ("similar form"), it draws attention to higher levels, strengthening expectancies on these levels relative to the lower tone-to-tone level.

Krumhansl (1997) empirically demonstrated the influences of expectancies on higher levels. Narmour composed a number of short melodies that varied in form; they were either "similar form," AA', or "dissimilar form," AB. They were arranged so that the implications at high and low levels would be opposite, depending on the form AA' or AB. In the experiment, the sequences were followed by continuation tones that listeners judged as to whether they were good continuations of the melody. The results confirmed the model's prediction that the strength of the implication-realization principles on the higher-level should be greater for similar form melodies, whereas the strength of the principles on the lower-level should be greater for dissimilar form melodies. This was true for all five basic melodic principles. Thus, form is one factor that might be taken into account in attempting statistical summaries of important non-adjacent relationships.

A second point concerning the implication-realization model is that the principles are intended to apply only at certain points in music when expectations are strong, not just all note-to-note intervals as were counted by Zivic et al. (2013). Thompson and Stainton (1998) analyzed successive intervals in folk melodies, which were divided into two classes: strongly implicative (when two sequential notes moved to a weaker metric position, decreased tonal stability, and decreased or unchanged duration), and closural (if two sequential notes moved to a stronger metric position, increased tonal stability, and increased or unchanged duration). Their Table 1 shows large effects for three of the principles: registral direction and intervallic difference were satisfied considerably more often in implicative than in closural situations, and the opposite was found for the principle of melodic closure. It may be that because Zivic et al. (2013) ignored these contextual differences, they found that the I-R principles accounted less well for the patterns of intervals in their corpus than the factors derived from their analysis.

Zivic et al. (2013) also conducted an analysis that examined whether there were trends over time in the degree to which the principles of the implicative-realization model were evident in their musical corpus. Although some trends were apparent, they found less clear distinctions between musical styles than the factors derived from their analysis. However, this would not be surprising in light of the theoretical nature of the proposed I-R principles. They are presumed to be similar to Gestalt principles in that they are considered innate and universal, so no stylistic shifts would necessarily be expected.

Nonetheless, cross-cultural studies suggest the principles may be differentially weighted in different styles. In fact, a study of melodic implications in the Sami yoiks (Krumhansl et al., 2000) uncovered an additional principle: preference for large, consonant intervals. This was found independently of whether or not the listeners were familiar with the style. Findings such as this argue against attempts to reduce the I-R model to fewer principles (e.g., Schellenberg, 1997). At a conceptual level, this cross-cultural experiment suggest that principles such as those in the I-R model might better be considered cognitively latent; they can be evoked to different degrees depending on the patterns present in the music, dynamically guiding expectations rather than being applied mechanically. Further studies with music outside the tonal-harmonic tradition may point to other latent principles employed in non-Western and contemporary music.

Finally, it should be mentioned that the principles themselves interact with one another. Under some conditions, groups of principles will converge on the same set of implied tones, whereas under other conditions they will not. Considering this, shifting attention to different levels, and variations in the strength of the implications over time leads to a much more dynamic concept than a mechanical application of fixed principles. Krumhansl (1997) summarized, "Music capitalizes on the interactions possible between these principles, which are especially complex when a number of different levels of patterning are present. This produces a dynamic, constantly changing engagement of cognitive processes by the music as it unfolds in time. Similar processes operate on metrical, harmonic, and rhythmic aspects of the music . . . and similar complexity may be found there as well." I will return later to the issue of the interactions between these different aspects of music.

The issue of hierarchical organization may perhaps become clearer when considering the tree structures of Lerdahl and Jackendoff (1983, henceforth GTTM) and Lerdahl (2001, henceforth TPS). In these trees, typically about half of the events are depicted in non-adjacent relationships. That is, they are subordinate in the tree to an event that is neither immediately preceding nor following it on the musical surface. Links between non-adjacent events are called in linguistics "long-distance dependencies." In what follows, I will primarily refer to the prologational trees of TPS because it is a fully quantitative theory that has been tested empirically with a variety of different contexts in different musical styles (Lerdahl & Krumhansl, 2007).

The model predicts tension using four components. The first component calculates distances between events (any chord in any key to any other chord in any other key). The distances take into account the required changes in the "basic pitch space," chord distances, and key distances. The latter two are represented geometrically in the "chordal-regional space." The second component is the prolongational tree that specifies which events are subordinate to which others in the event hierarchy. What is called "global" tension is computed as the sum of the distances from any event to all the events that are superordinate to it in the tree, until the most important or stable event in the passage is reached. Predicted tension will generally be high (depending on the specific distances) if the event is deeply embedded in the tree, subordinate to many other events. The global tension computed in this way is added together with two other components of the model, attraction (which is calculated between successive events) and dissonance, giving the final predicted tension value. These quantified predictions accounted precisely for judged tension in the experiments (Lerdahl & Krumhansl, 2007).

In one extended analysis, Lerdahl and Krumhansl (2007) compared sequential and hierarchical predictions and demonstrated that computing tension according to the prolongational tree provided a far better fit than computing tension from each event to the next. Moreover, it should be stressed that the match between prediction and data for every single event in all of the examples tested is a confirmation of the hierarchical tree model. This is because the local tension prediction is derived quantitatively for each event in each sequence from the prolongational tree. This precise quantitative fit is strong confirmation of the specific structural relationships represented in the tree. At a general level, it may be that the long-distance dependencies are especially significant both psychologically and theoretically because they are potent enough to overcome the tendency toward grouping by temporal adjacency.

More recently, Rohrmeier (2011) proposed a hierarchical model, called the generative syntax model. Briefly, it derives a hierarchy by identifying embedded tonic, dominant, and subdominant regions in the spirit of Hugo Riemann. Koelsch, Rohrmeier, Torrescuso, and Jentschke (2013) tested the model using contrasting versions of two Bach chorales: one in which the regular hierarchical structure between the first and second phrases was present, and one in which it was made irregular by altering the first phrase while keeping the second phrase intact. They found an effect on the final chord as predicted. They argued that because the EEG activations were similar on the second to last chord in the two versions (thus "proving" the null hypothesis), the effect was the result of changing the long-distance dependency between the first and second phrase and not a local effect.

It is important to be clear about the nature of hierarchical models and their empirical tests because of their broader significance for commonalities between music and language. Hauser, Chomsky, and Fitch (2002) claim that recursion in language (embedding phrases within phrases) is the only computational mechanism in language that is unique to humans. "Natural languages go beyond purely local structure by including a capacity for recursive embedding of phrases within phrases, which can lead to statistical regularities that are separated by an arbitrary number of words or phrases. Such long-distance, hierarchical relationships are found in all natural languages for which, at a minimum, a "phrase-structure grammar" is necessary. It is a foundational observation of modern generative linguistics that, to capture a natural language, a grammar must include such capabilities." (p. 1569). (See, however, Frank, Bod, & Christiansen, 2012, for a contrasting position arguing that knowledge of sequential sentence structure gained through statistical learning is largely sufficient.)

Lerdahl and Krumhansl (2007) noted some differences in detail between traditional linguistic trees and the prolongational trees of TPS. In traditional linguistic grammars, for example, the trees have labeled links, such as for NP (noun phrase) and VP (verb phrase), which is more like the tonic, dominant, and subdominant labels in Rohrmeier's (2011) generative syntax model. However, it is interesting to observe the trend toward label-less links and dependency relationships in linguistic trees (e.g., Osborne, Putman, & Gross, 2011) in the minimalist program starting in the 1990s (e.g., Chomsky, 1995). Thus, the linguistic trees resemble increasingly the prolongation trees of GTTM and TPS.

To further explore possible music-language parallels, Erin Isbilen and I recently conducted a pilot language-tension experiment, in which each successively presented word in sentences from Grosjean, Grosjean, and Lane (1979) was rated for tension. The tension ratings correlated negatively and strongly with their complexity measure, which was the number of hierarchical constituents that were closed off by the word. This would be analogous (inversely) to counting the number of superordinate nodes of an event in the tree. The tension ratings also correlated with the pause durations when the sentences were read aloud, similar to the more exaggerated rubato found in music performance at the end of major sections, as represented in the time-span reduction trees of GTTM, for example (Todd, 1985). These findings suggest further language-tension studies may prove fruitful for exploring more deeply parallels between music and language.

Returning now to general issues in statistical treatments of music, a second limitation of most statistical treatments of music is that they are unidimensional, usually focusing on pitch without considering temporal factors (meter or duration) or focusing on temporal factors without considering pitch (tones or chords). The empirical literature, as reviewed by Prince and Schmuckler (2014), is remarkably inconsistent on the extent and nature of perceptual interactions between pitch and time. It is also notable that theirs is the first relatively large statistical corpus analysis that considers both pitch and time, specifically metrical and tonal hierarchies in a selection of tonal-harmonic music. (Järvinen, 1995, had done a similar analysis on a smaller corpus of jazz improvisations, and Temperley, 2014, included metrical information in a recent analysis of repeated intervallic patterns.) They determined the frequency of occurrence of each of the 12 pitch classes at all possible temporal positions in a variety of different meters. They found that composers do indeed use tonally stable pitches at points of high metric stability. The tonal hierarchy was clearest on the downbeats of measures, followed by half measures, and so on. This demonstrates that there is indeed an interaction between these dimensions and that statistical analyses of pitch might benefit from considering temporal factors and vice versa.

A machine learning study conducted a number of years ago with David Rand approached the question of how pitch and time are related using a different approach. We tested whether an unsupervised learning algorithm (ADIOS, Solan, Horn, Ruppin, & Edelman, 2005) is affected by randomizing the durations in melodies taken from the classical collection in Themefinder.com. The learning algorithm abstracts two kinds of information: patterns and equivalence classes. A *pattern* is a similarly structured sequence of elements that recurs in the corpus. An *equivalence class* is a set of elements that occur in the same position in a pattern. Patterns can be included in equivalence classes, and equivalence classes can be included in patterns. (See Solan et al., 2005, for technical details.) The melodies were encoded in Kern format.

It should be emphasized that the model treats the Kern elements, the duration/pitch pairs, as single units. As far as it is concerned, 4G is just as similar to 8d+ as it is to 8G or 4A. In other words, it has no way of knowing that duration/pitch pairs have similar durations or similar pitches, what pitches are in diatonic scales, or what pitches are proximal.

The primary manipulation was creating what we called *scrambled* melodies in which the order of the pitches was unchanged but the durations were randomly reassigned to the pitches. If there are regular relationships between pitch and duration, for example if more stable pitches are played for longer durations, then scrambling the durations in this way might interfere with what the model extracts from the corpus. Two models were trained, one on 300 original melodies in C major and one on 300 original melodies in A minor. Two other models were trained on the same sets of melodies in which the durations were randomly reassigned to the pitches, the scrambled melodies. Thus, we had four models trained on different corpora: C major original melodies, C major scrambled melodies, A minor original melodies, and A minor scrambled melodies. The question was whether the algorithm is less able to abstract interpretable regularities in melodic structure when the durations were randomly reassigned to the pitches.

We found that the models trained on the original (unscrambled) melodies outperformed those trained on the scrambled melodies in a number of ways. The models trained on the original melodies extracted more patterns and equivalence classes than those trained on the melodies with scrambled durations. They were also more coherent, tending to group together different pitches of the same duration, and proximal diatonic and chromatic scale tones. The contents of the patterns and equivalence classes in the models trained on the original melodies also correlated more strongly with the probe tone ratings (Krumhansl & Kessler, 1982). They were better at distinguishing between novel C major and A minor melodies.

Interestingly, when listeners in an accompanying perceptual experiment were asked to distinguish between major and minor melodies, their performance was not strongly affected by scrambling the durations. The participants may have been able to filter out the durational variation when making the mode judgments. However, when they also judged how well structured the melodies were, this was affected by randomly reassigning the durations, suggesting they are aware of the typical relationships between pitch and duration.

This machine learning study and the corpus analysis performed by Prince and Schmuckler (2014) both suggest that some relatively straightforward research might profitably be done to determine which parameters of music are not independent of one another. The most likely candidates are those concerning pitch and time. In addition to the obvious candidates of tonal and metrical hierarchies and tone duration, other parameters might include pitch range, tone density, and consonance. Even timbre and dynamics are possible candidates, although not all parameters of music may have psychological correspondences. "... it does not follow from what seems to be the statistical character of first and second-order harmonic relationships that all parameters are statistical or that all hierarchic levels are so." Meyer (1967, p. 259). The results of statistical studies considering multiple parameters in combination might help suggest further experimental tests that would serve to clarify exactly what musical parameters interact in perception and cognition, and under what conditions.

The third, and final issue concerning statistical approaches to music is the assumption that probabilities are relatively uniform throughout a piece of music. To quote from Meyer (1967), "It is a mistake to suppose that probability remains relatively constant throughout musical works. Quite the contrary. Some parts of a work tend to adhere much more closely to the normative and probable than do other parts.... Thus serious statistical and methodological errors arise if probabilities are computed on the basis of a total "average" frequency over the entire piece. Subsystems must be analyzed within the larger probability system." (p. 19).

Moreover, Meyer (1967) offers an interesting psychological reason for why probabilities should vary over the course of a piece: "... music is not a natural system. It is man-made and man-controlled. And it is able to combat the tendency toward the tedium of maximum certainty through the *designed uncertainty* introduced by the composer... And there may well be a difference between the probabilities of systematic origin [objective statistics] and those introduced by design. That is, we should expect designed deviations, delays, and ambiguities to be introduced as systematic probability increases – as the pattern approaches completion." (p. 15). In line with this, experiments measuring musical tension tend to find a pattern in which tension increases relatively gradually and then decrease rapidly as the phrase concludes. Krumhansl (1996) proposed a tension schema in which a section begins with new ideas, played at a neutral tempo and at a low level of tension. Then, through the phrase, tension increases and there are increased dynamics and tone densities. The tempo slows near the end of the phrase and tension decreases rapidly as the section ends. This kind of tension profile has been found at both local and global levels for dance as well as music (Krumhansl & Schenck, 1997) and may generalize to events in film and literature, and possibly more naturally occurring processes.

This pattern of gradually increasing tension followed by a more rapid decrease to stability suggests that it might be informative to conditionalize statistical analyses on the relative position of events within phrases. For example, one might separately focus on the first half of the phrase, the next third where tension tends to peak, and then the remaining end of the phrase. Rather different statistical processes may occur during these segments. Alternatively, it might be possible to derive tension profiles by measuring the degree to which local statistics deviate from the overall norm.

I would like to end with a few final comments about statistical learning and music. Despite considerable evidence that listeners are sensitive to distributional information in music, there may be limits to what can be learned statistically. One example that comes to mind is a probe tone experiment reported in Krumhansl, Sandell, and Sergeant (1987). The prime forms of the 12-tone series from Schoenberg's *Wind Quinte*t and *String Quartet*, *No. 4.* were probed after the first 3 tones, then after the first 6, then after the first 9, and finally after the complete 12-tone series. The procedure was then repeated and one question was whether, on the second time through, listeners gave higher ratings to the next tone that actually follows in the series. Despite the number of repetitions (36 times for the transition between the 3rd and 4th tones, for example), the musically trained participants who were unfamiliar with atonal theory actually gave lower ratings to the correct following tone. Their responses were dominated by the influence of local tonal expectations. The participants familiar with atonal theory gave slightly higher ratings, but appeared to do so by reversing local tonal expectation. It is not the case, however, that participants learned nothing about the series. In two other experiments (one using neutral presentations of the series and one using actual excerpts from the pieces), all participants distinguished between the two compositions in the four forms of the series: prime, retrograde, inversion, and retrograde inversion. The different intervallic content of the two series (such as the relative presence of whole tones) may have enabled them to make the distinction.

At the other extreme perhaps is the remarkable sensitivity to a novel style exhibited in an experiment using Messiaen's *Mode de valeurs et d'intensité* (Krumhansl, 1991). This piece is written according to a unique set of constraints that rigidly couples values of pitch (chroma and octave), duration, and dynamics. The participants heard only the first half of the piece, and then were presented with excerpts that they were to judge as to whether or not they might have come from the piece (either from the part they had heard or from the remainder). Even on the first block of trials, they were able to recognize segments they had heard and also accurately accept segments from the part of the piece they had not heard, while rejecting transformations that decoupled the global correlations between pitch height and duration, dynamics, and interval size. Performance was accurate and did not improve with repeated hearings of the music over the next 5 blocks of trials. Most accounts of statistical learning would not predict this immediate degree of sensitivity to the complex coupling across multiple dimensions.

In sum, I have considered three problems with unidimensional, sequential statistics. First, music is hierarchical and sequential statistics are not sensitive to significant relationships that may exist between non-adjacent events. Second, the various dimensions of music may interact, not only statistically, but also in perception and cognition. Finally, probabilities do not necessarily remain constant throughout a composition, and listeners may expect varying degrees of deviation from normative probabilities at different points in phrases and different points in the overall musical form. However, measurement is essential in psychological research, and probabilities are a convenient way to capture generalities.

The last three decades have witnessed exponential growth in empirical research on how music is perceived, organized, remembered, and performed. The results of large-scale statistical studies may well uncover interesting and significant patterns not known previously from the experimental research. Such findings in turn might stimulate studies aimed at determining whether the statistical patterns observed have subjective correlates and what psychological processes are involved. Without this, we only know something about how music is structured, but not why it is structured in that way. In a complementary way, results of empirical studies of the human mental processes engaged by music might potentially lead to the development of novel statistical techniques that are more informed by an understanding of the complex and dynamic nature of the musical experience.

Author Note

The comments of Davy Temperley and Fred Lerdahl on an earlier version of this manuscript are gratefully acknowledged.

Correspondence concerning this article should be addressed to Carol Lynne Krumhansl, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY, 14853, USA. E-mail: clk4@cornell.edu

References

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

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

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of North India. *Journal of Experimental Psychology: General, 113,* 394-412.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Cuddy, L. L. (1997). Tonal relations. In I. Deliège & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 329-352). Hove, England: Taylor and Frances.

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

Eerola, T., Järvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception, 18*, 275-296.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, 57, 94-107.

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences, 279*, 4522-4531.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology, 11,* 58-81.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569-1579.

Hyer, B. (2003). Tonality. *Grove Music Online* (L. Macy, Ed.). Accessed 23 October 2003, http://ezproxy.library.cornell.edu:2691

Järvinen, T. (1995). Tonal hierarchies in jazz improvisations. *Music Perception, 12*, 415-437.

Koelsch, S., & Rohrmeier, M., Torrescuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences, 110,* 15443-15448.

Koffka, K. (1935). *Principles of Gestalt psychology.* New York: Harcourt, Brace, and World.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85,* 445-463.

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11*, 346-384.

Krumhansl, C. L. (1985). Perceiving tonal structure in music. *American Scientist, 73*, 371-378.

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

Krumhansl, C. L. (1991). Memory for musical surface. *Memory and Cognition, 19*, 401-411.

Krumhansl, C. L. (1995). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum, 17*, 53-80.

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. (1997). Effects of perceptual organization and musical form on melodic expectancies. In M. Leman (Ed.), *Music, Gestalt, and computing: Studies in cognitive and systematic musicology.* Berlin: Springer-Verlag.

Krumhansl, C. L., Bharucha, J. J., & Kessler, E. J. (1982). Perceived harmonic structure of chords in three related musical keys. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 24-36.

Krumhansl, C. L., Bharucha, J. J., & Castellano, M. A. (1982). Key distance effects on perceived harmonic structure in music. *Perception and Psychophysics, 32,* 96-108.

Krumhansl, C. L., & Cuddy, L. L. (2010). A theory of tonal hierarchies in music. In M. R. Jones (Ed.), *Springer handbook of auditory research* (pp. 51-87). New York: Springer.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial map of musical keys. *Psychological Review, 89,* 334-368.

Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic expectancy in Finnish folk hymns: Convergence of behavioral, statistical, and computational approaches. *Music Perception, 17*, 151-196.

Krumhansl, C. L., Sandell, G. J., & Sergeant, D. C. (1987). The perception of tone hierarchies and mirror forms in twelve-tone serial music. *Music Perception, 5,* 67-113.

Krumhansl, C. L., & Schenck, D. L. (1997). Can dance reflect the structural and expressive qualities of music? A perceptual experiment on Balanchine's choreography of Mozart's *Divertimento No. 15. Musicae Scientiae, 1,* 63-85.

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within the diatonic context. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 579-594.

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

Large, E. W. (2011). Dynamics of musical tonality. In R. Huys & V. Jirsa (Eds.), *Nonlinear dynamics in human behavior* (pp. 193-211). New York: Springer.

Large, E. W., Kim, J. C., Flaig, N. K., Bharucha, J. J., & Krumhansl, C. L. (in press). A neurodynamic account of musical tonality. *Music Perception.*

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception, 17*, 481-509.

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.

Lerdahl, F., & Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception, 24*, 329-366.

Luce, R. D., & Krumhansl, C. L. (1988) Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., pp. 3-74). New York: John Wiley & Sons.

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

Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism, 15*, 412-424.

Meyer, L. B. (1967). *Music, the arts, and ideas: Patterns and predictions in twentieth-century culture.* Chicago, IL: University of Chicago Press.

Meyer, L. B. (1989). *Style and music: Theory, history, and ideology.* Philadelphia, PA: University of Pennsylvania Press.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model.* Chicago, IL: University of Chicago Press.

Osborne, T., Putman, M., & Gross, T. M. (2011). Bare phrase structure, label-less trees, and specifier-less syntax. Is minimalism becoming a dependency grammar? *The Linguistic Review, 28,* 315-364.

Oram, N., & Cuddy, L. L. (1995). Responsiveness of Western adults to pitch-distributional information in melodic sequences. *Psychological Research, 57*, 103-118.

Prince, J. B., & Schmuckler, M. A. (2014). The tonal-metric hierarchy: A corpus analysis. *Music Perception, 31,* 254-270.

Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music, 5,* 35-53.

Rohrmeier, M., & Cross, I. (2013). Artificial grammar learning of melody is constrained by melodic inconsistency: Narmour's principles affect melodic learning. *PLoS ONE, 8,* e66174. doi: 10.1371/journal.pone.0066174

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7, 532-547.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.

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

Schoenberg, A. (1978). *Theory of harmony* (R. E. Carter, Trans.). Berkeley, CA: University of California Press. (Original work published 1906)

Shaffer, L. H. (2007). Critical notice: Towards a cognitive psychology of music. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 44*, 177-183.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. In *Proceedings of the National Academy of Sciences, 102,* 11629-11634.

Temperley, D. (2014). Information flow and repetition in music. *Journal of Music Theory, 58*, 155-178.

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

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

Todd, N. (1985). A model of expressive timing in tonal music. *Music Perception, 3,* 33-58.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.

Wertheimer, M. (1938). Gestalt theory. In W. D. Ellis (Ed. & Trans.), *A sourcebook of Gestalt psychology* (pp. 1-11). London: Routledge and Kegan Paul.

Zivic, P. H. R., Shifres, F., Cecchi, G. A. (2013). Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences, 110*, 10034-10038.