Order Number 9119658

Simulation and evaluation of a model of meter perception

Miller, Benjamin O., Ph.D.

City University of New York, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A

Simulation and Evaluation of a Model of Meter Perception

by

Benjamin O. Miller

A dissertation submitted to the Graduate Faculty in
Psychology in partial fulfillment of the requirements
for the degree of Doctor of Philosophy,
The City University of New York.

1991

This manuscript has been read and accepted for the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Jan 18, 1991
Date

Don L Scarborough
Chair of Examining Committee

January 24, 1991
Date

Herbert D. Saltzstein
Executive Officer

Supervisory Committee: Don L. Scarborough

Israel Abramov

Neil A. Macmillan

The City University of New York

Abstract

Simulation and Evaluation of a Model of Meter Perception

by

Benjamin O. Miller

Adviser: Professor Don L. Scarborough

A model of duration-based meter perception was developed and implemented as a computer simulation. The model, BEATS, produces a hierarchical representation of meter within the first few notes of a piece of music. BEATS was evaluated as a model of human meter perception in four experiments. The first three used the metric probe technique, in which the subject hears a very short rhythm, which is hypothesized to induce a particular meter, followed by a single probe note. The delay of the probe places it at one of several positions in the hypothesized metric hierarchy. The pattern of subjects' responses to probes in different positions served as an indication of perceived metric structure. Each experiment involved a different response: rating the probe for how well it completed the rhythm; identifying the note value corresponding to the probe's delay; and reproducing the rhythm and probe by tapping. Results from these experiments suggest that listeners do represent meter hierarchically and that, like BEATS, they can extract meter

quickly and from durations alone. The fourth experiment compared BEATS analyses of longer musical rhythms with the metric structure perceived by subjects. Subjects' task was to tap the beat of the rhythm being played. By tapping with both hands subjects revealed two levels of the perceived metric hierarchy, one level to each hand. BEATS correctly predicted 40% of subjects' responses, both correct and incorrect. In addition, individual subjects tended to produce more than a single interpretation of individual stimuli, which is inconsistent with the fact that BEATS produces a single analysis of a given piece. This may be resolved by considering that a) listeners normally extract meter not from duration alone but from several kinds of cues (including duration) at once; and that b) when other cues are unavailable duration cues may be inadequate or ambiguous.

Acknowledgements

This research was supported in part by a National Science Foundation Graduate Fellowship awarded to the author and by City University of New York Faculty Research Awards to Prof. Don L. Scarborough and to Prof. Jacqueline A. Jones.

For Sue

Table of Contents

List of Tables

NOTE: The first digit of a table number refers to an experiment.

List of Figures

NOTE: The first digit of a figure number refers to an experiment. Figures not associated with a particular experiment have a first digit of zero.

I. Introduction

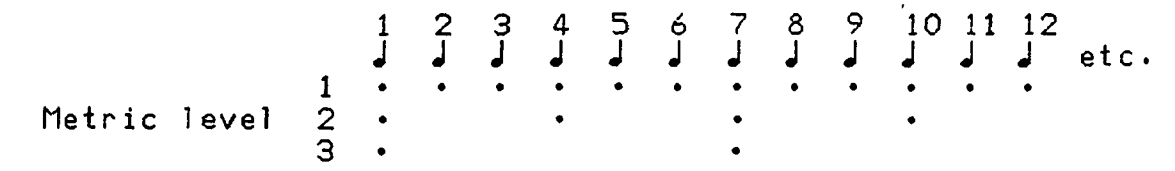
The Problem

You turn on the radio. A symphony you have never heard is being played. A moment or two later, perhaps without thinking about it, you begin conducting the invisible orchestra. What has happened here?

What follows is concerned with several aspects of the above scenario and others like it. First, what are the temporal characteristics of this conducting (toe-tapping, finger-snapping, table-rapping, etc.) response? What is the nature of the mental representation of these time-keeping activities? What is the relation between this representation and the musical stimulus? What aspects of music are relevant to this problem? Most important, how does the listener generate this representation as he/she listens to a piece of music?

Metric Structure

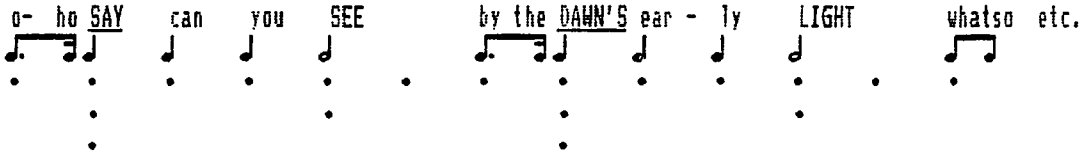
People perceive patterns in temporal events even in the absence of any physical cues. For example, we hear a sequence of identical, equally spaced tones as being grouped by twos or possibly threes (Fraisse, 1982), with the first tone of each group accented, as shown below.



In this notation, introduced by Lerdahl & Jackendoff (1983), each row of dots is a meter, a regular series of perceived pulses, or beats, that mark off equal temporal intervals. Each meter is a level in a metric hierarchy. At each successive level of the hierarchy the interval between beats is double or triple that of the previous level. When beats coincide, as at locations 1, 4, 7 and 10 above, they together constitute a single beat that marks off intervals at each of the several metric levels, and as such are stronger beats. The perceptual aspect of the above metric structure can be reproduced by tapping with the number of fingers indicated by the number of dots in a column at each note (3 1 1 2 1 1 etc.), or tapping with a force proportional to the highest level at which a given beat measures off an interval, or waving one's arm in a roughly triangular pattern, and so on. Thus metric structure is more than an organization of meters; it is a hierarchy of accents as well.

Meters and metric hierarchies are mental representations, but they interact with music in a complex way. While a sequence like that above can be perceptually organized in several different ways, more interesting

pieces are generally consistent only with particular metric structures, as illustrated below. Certain points in the music are stressed, as indicated by capitals (slight stress) and by underlined capitals (stronger stress). In the first example, a metric structure that groups ♩s by three (i.e. a 3/4 time signature) is consistent with the spacing of stressed notes in the song.



However, the same piece is inconsistent with a different metric structure (below) that groups ♩s by two (i.e. a 4/4 time signature) because some strong beats coincide with unstressed notes (or with no note) in the song.

In the above examples, the first meter is better because it accurately reflects the regular recurrence of stresses in the music. This interaction between metric structure, which predicts stresses, and the music itself is rhythm. For example, syncopated rhythms occur when perceived musical accents are heard at relatively weak positions in the metric structure. Thus, musical events are heard within a framework established by the metric structure. On

the other hand, the music itself must also guide the listener in establishing that metric framework in the first place.

The question addressed by this research, then, is this: How does the listener discover a metric structure? In particular, to what extent does the sequence of durations contained in a piece of music provide the listener with the means to extract an appropriate mental representation of the music's temporal regularities? Concentrating on the role of duration in this way is not meant to imply that other musical dimensions (e.g. melody, harmony, lyrics) are not relevant to the extraction of meter or to the rhythmic characteristics of music in general. However, there are three justifications for limiting the present research to the extraction of meter from duration. First, given the perceptual richness of music and the relative theoretical and empirical poverty of music psychology, it is necessary to begin with fairly simple stimuli. While it is not clear how or to what extent duration contributes to meter extraction, it is nevertheless the most obvious place to begin, and it is clear from preliminary results, discussed below, that listeners can and do find metric structure in duration sequences. Second, by understanding when and how duration fails to yield a clear or correct perception of meter, we will have learned something about what to look

for in examining the contribution of other musical dimensions to meter perception. Third, this approach allows an indirect evaluation of the suggestion that the metric and melodic dimensions of music are perceived independently and represented by separate hierarchies (Palmer & Krumhansl, 1987a, b). If this is the case we should expect that the series of durations in music contains all the information a listener needs to extract the metric hierarchy. To the extent that this does not happen in other experimental paradigms, we will have to question the notion of perceptual independence of meter and melody.

Perception of Hierarchies

Following Lashley's (1951) demonstration of the shortcomings of associative chains, hierarchic structures have come to dominate models of representation of temporal sequences (e.g. Simon & Sumner, 1968; Restle, 1970; Jones, 1978; Deutsch & Feroe, 1981). Hierarchies embody both order and structural relations, and are often represented by trees or nested formulas. The set of possible events in a class of sequences is designated as an alphabet. Some alphabets used in representation of music might be the notes of the chromatic scale, major and minor diatonic scales, triads, arpeggios, and so on. In addition to alphabets there are operators, which when applied to an

element or sub-sequence, with reference to some alphabet, yield another element or sub-sequence. Formulaic representations of sequences consist of nested operations and specifications of alphabets. The idea that music is represented by such structures accounts nicely for the fact that recognition of melodies is not affected by transposition or, within limits, by tempo changes. At a more detailed level, the complexity of a formula can predict how accurately the corresponding musical sequence is perceived (Jones, Maser & Kidd, 1978).

Evidence of hierarchies in the perception of temporal sequences comes from Perkins (1974), who played long isochronous series of taps, with every fourth tap accented, and asked subjects to reproduce the series. Reproductions were accurate with respect to the position of the last tap relative to the nearest accented tap, but inaccurate with respect to the number of four-tap groups in the series. Perkins suggests that each group of four is encoded hierarchically but that the number of groups of four is not encoded. One way of interpreting this is that there is nothing in the stimulus that encourages grouping the groups of four into higher level groups. Similarly, Povel (1981) found that sequences of tones whose durations stood in ratios such as 1:1:3 were less accurately reproduced than those with ratios such as 1:1:1:3. The latter sequence can

be divided into sub-sequences of equal duration (1:1:1 and 3) but the former cannot. Such division constitutes a balanced hierarchical representation, and Povel's results suggest that sequences of durations are stored in this way when possible. This may help explain the fact that in music 1:1 and 2:1 are by far the most common ratios between durations of adjacent notes. For example, the relative durations 1, 1 and 2 can be arranged in three ratios, two of which (2:1:1 and 1:1:2) are balanced in the sense described above. But as the number of different relative durations grows, the number of balanced ratios that can be found in a sequence of a given length becomes proportionally smaller. By using a small number of relative durations standing in small ratios to one another, a piece of music facilitates its representation in terms of balanced trees.

In addition to being more accurately reproduced, there is evidence that sequences that can be represented hierarchically are more easily learned and better remembered. For example, Deutsch (1980) found that transcription of brief melodies was more accurate when melodies were hierarchic than when they were not. Likewise, Restle (1972) found that serial patterns of lights were more quickly learned when they were temporally segmented in a way that complemented their serial

structure. Sturges and Martin (1974) had similar results using short binary patterns. Simon & Kotovsky (1963) found that the number of subjects who discerned the hierarchical pattern in a letter series completion task was predicted by the complexity of the pattern.

Other evidence for hierarchic representation comes from studies of click localization. It has been shown that a click superimposed on a spoken sentence tends to migrate perceptually toward the nearest syntactic break. This has been interpreted as evidence of the psychological reality of linguistic units (Fodor & Bever, 1965), though the phenomenon can be accounted for at least as well by an attentional hypothesis (Reber & Anderson, 1970; Reber, 1973). Accepting the linguistic interpretation, Gregory (1978) investigated the intuition that what holds for syntactic hierarchies should hold for musical hierarchies as well. If the click migration observed in the linguistic domain is evidence of hierarchical representation, then the same phenomenon in the musical domain would be evidence for the hierarchical representation of musical sequences. Gregory's results on the whole support his hypothesis, and Stoffer (1985) has replicated and extended Gregory's results.

Deutsch & Feroe (1981) point out that hierarchically structured stimuli can be encoded more parsimoniously than

non-hierarchic sequences. While this economy varies from one scheme of alphabet-based coding to another, the fact remains that the more regular a sequence is, the more concisely it can be encoded. The important question is whether the alphabets and operators used by a particular coding scheme can be plausibly attributed to a human listener.

Martin (1972) describes a different advantage for hierarchic representations. If important musical events tend to coincide with strong beats in the metric hierarchy, then a mental representation of the metric hierarchy will allow the listener to predict the occurrence of such events in time. These predictions may allow the listener to allocate attention more efficiently in the interims, much as a fixed interval reinforcement schedule allows a pigeon to allocate pecks more efficiently.

Lerdahl & Jackendoff's Generative Theory

Martin's observation is one of the implicit cornerstones of Lerdahl & Jackendoff's (1983) A Generative Theory of Tonal Music (henceforth GTTM), in which it is assumed that our perceptual processes strive to produce an interpretation in which parallel structures (e.g. melodic and rhythmic structures) complement each other as much as possible. GTTM attempts to formalize the intuitions of a human listener regarding classical Western tonal music.

The theory does this by means of four stages of analysis, each embodied in two sets of rules. Well-formedness rules are analogous to grammatical rules, in that they specify legal structures within a stage, and preference rules correspond to laws of perceptual organization.

One of Lerdahl & Jackendoff's key innovations in GTTM is the explicit partitioning of rhythm into two independent hierarchic components: metric structure and grouping structure. The metric analysis stage of the theory yields a hierarchical representation of metric structure which conforms to traditional intuitions about meter and accent. The grouping stage yields another hierarchy, reflecting intuitions about musical phrases. Grouping and metric analysis are largely independent, and each can, to a large extent, be carried out without the other. In everyday usage, "rhythm" sometimes refers to the sequence of note durations in a piece of music, sometimes to the way those notes are grouped, objectively or perceptually, and sometimes to meter. In GTTM, rhythm is the interaction, in the listener's mind, of all of these aspects of a piece of music. This more complex notion of rhythm is possible only when the grouping/meter distinction is made.

Grouping, in GTTM, is the division of a musical passage into motives, themes, phrases, and the like, up to and including the piece itself. Grouping well-formedness

rules say, for example, that a group may be partitioned (exhaustively) into subgroups, with each subgroup being wholly contained by the larger group. Grouping preference rules prescribe where to find group boundaries. For example, for any four-note sequence, a boundary will be found between the middle two notes if the difference between those notes is larger with respect to scale distance, loudness, articulation or duration than the differences between the first two and the last two notes.

The metric hierarchy is largely independent of the grouping hierarchy, and its structure is constrained by metric well-formedness rules, which require, for example, that every note onset must be represented by a beat at the smallest metric level and, conversely, that every beat at a given level must also be a beat at all smaller levels. The metric preference rules prescribe how to decide among alternative well-formed hierarchies. The most basic of these rules gives preference to a metric hierarchy in which strong beats (i.e. beats at higher metric levels) coincide with stressed notes.

A difficulty for any hierarchic conception of meter is specifying which level(s) of the hierarchy correspond(s) to a listener's perception. GTTM's meter rules allow the analyst to construct a hierarchy with an indefinite number of metric levels, from the shortest note duration to spans

of many measures. While it seems unlikely that all these levels are simultaneously present and equally salient in the average listener's experience, it is not clear how to determine which levels are most important. For one thing, individual differences are likely to confound any general solution. A musically trained listener may hear more metric levels simultaneously than a naive listener, and the demands of the situation (e.g. dancing vs. conducting) may also influence the listener's metric experience. In addition, tempo is certain to be an important factor: as a given piece is played faster, higher metric levels (i.e. levels corresponding to larger note values) become more apparent. Lerdahl and Jackendoff have captured this intuition in their description of the tactus, or most salient metric level, as that level at which beats pass by neither too quickly nor too slowly.

Models of Rhythm Perception

GTTM comprises some fairly detailed intuitions about the mental representations that result from the perception of music. The perceptual and cognitive processes involved in generating these representations are, for the most part, beyond the scope of GTTM. A number of workers have begun taking a psychological approach to GTTM. Deliege (1987) has conducted experiments to determine whether GTTM's grouping rules are reasonable representations of human

perception of grouping. Her work seems to support GTTM. Tenney & Polansky (1978, 1980) have written a simulation based on a set of temporal gestalt rules that are similar in spirit to GTTM's grouping rules. A weakness of this program is that its many parameters apparently must be adjusted differently for different scores.

A recent model by Rosenthal (1989) is explicitly based on the grouping rules of GTTM. Input is duration-only but includes bar lines, and the program moves in one pass through the score. Grouping rules are used to identify candidate groups, and the model includes rules for choosing between competing candidates. As groups of notes are identified by these methods, they are simultaneously grouped in a hierarchy of groups. Because groups (at any level) are not necessarily the same size, the mapping between the hierarchy of groups and the metric hierarchy (which is not produced by Rosenthal's program) is not always perfect. On the other hand, in GTTM grouping is a function not only of duration but of pitch as well. Considering that Rosenthal's groups are determined only by those grouping rules that deal with duration, it is not clear just what the resulting grouping hierarchy corresponds to in human experience.

Other models have taken different approaches. Simon's (1968) multi-pass LISTENER program groups the note

durations it uses as input into note groups of equal duration. From this it identifies repeating phrases that may span several measures. LISTENER does not distinguish between meter and rhythm, and cannot deal with initial upbeats or syncopation. Mont-Reynaud & Goldstein (1985) have developed more sophisticated algorithms that recursively discover patterns that occur more than once in a sequence. They have also developed a method for identifying two rhythmic patterns as having a common ancestor, i.e. being derivable, via a context-free grammar, from the same basic rhythm. Longuet-Higgins & Steedman (1971) developed a note-by-note parser that adopts the first note value as the basic metric unit and adjusts it based on subsequent note values. Among other things, the parser cannot handle passages of notes of equal duration. A later program by Steedman (1977), using the output of the Longuet-Higgins & Steedman program, makes a second pass through the score, considering not only note values but melodic repetition as well, assuming (as do Simon (1968) and Collard, Vos & Leeuwenberg (1981)) that size and separation of a melodic figure and its repeat reflect metric structure.

A program by Longuet-Higgins & Lee (1982) returns to a time-only orientation as well as a note-by-note approach. The program takes a list of note values (onset to onset) as

input, and uses four production rules (described in detail below) to generate a metric unit and the location of bar lines as output. This program is a major improvement over earlier programs (mentioned above) in that it handles syncopation and initial upbeats and produces a more detailed representation of meter. However, since it analyzes until the metric unit reaches a maximum size of one whole note and then stops, it cannot detect changes in meter in the middle of a piece, nor does it produce an analysis of the whole piece.

Grid theory (Povel, 1984) is a very different approach to meter extraction. Input is duration-only, but the program has access to all the input throughout its operation. A grid is analogous to the ticks of a metronome. The goal is to find the metronome rate and the placement of ticks (phase) that best fit the music. In general, a grid fits to the extent that its ticks coincide with notes (particularly accented notes) and do not fall in empty intervals. Povel has written a computer simulation based on a set of rules that quantify the fit of various possible grids (Povel & Essens, 1985). A shortcoming of the grid approach is that it identifies only the single best-fitting metric unit and therefore cannot detect a meter change.

Each of the above models produces as its output some aspect of a piece's meter or rhythm. Simon's (1968) LISTENER and Rosenthal's program identify rhythmic groups or phrases. The programs by Longuet-Higgins, Lee, and Steedman identify the time signature and place bar lines, with some additional information about the grouping structure within measures. Povel's program identifies a single 'best' metric unit, though it is not clear whether we are to consider this unit as the tactus or some other aspect of the listener's experience. While all of these are interesting and impressive achievements from a musical point of view, all have flaws as psychological models. The psychological issues that distinguish these models from one another and from the model to be presented here are the following:

1) Is the model's access to input limited to a single, left-to-right pass through the score, or does it either make several passes or have unlimited, random access to the score? The first approach seems more psychologically sound than the others because the limited capacity of working memory constrains the range of plausible models. In music, the limit is about six elements (Fraisse, 1982) or five seconds (Dowling & Harwood, 1986), and the same limits apply to verbal material (Turner & Poppel, 1983). While it may be argued that the human listener has random

access to mental representations derived from the stimulus, in the form of long-term memory for heard music, this is not the same as random (or repeated) access to the stimulus itself.

2) What information is available in the input? A psychological model should not use information that is not available to a human listener. Rosenthal's use of bar lines in his grouping rules may be an example of this. Bar lines can of course be inferred (with less-than-perfect accuracy) from metric structure or from the spacing of accented notes (which may themselves be inferred less than perfectly from durations), but there is no indication that this is in fact what Rosenthal's program does.

3) What is produced as output? A model's output should be easily interpretable in terms of human experience of music. Programs that produce time signatures or, better, bar lines, such as those of Longuet-Higgins, Lee and Steedman, would satisfy this requirement. LISTENER and Rosenthal's program produce representations that, while they may be psychologically valid, are harder to compare with the listener's experience.

4) Can the model handle initial upbeats, syncopation, and ambiguity? These seem to be the highest hurdles for any model of meter or rhythm perception. Grid theory ignores these issues, while Longuet-Higgins & Lee (1982) have paid

close attention to the first two. Lerdahl & Jackendoff (1983) suggest that ambiguity can be thought of in terms of conflicting well-formed hierarchies, but no model incorporates this notion.

II. The BEATS Model

Introduction

I have developed a new model, known as BEATS, as an attempt to remedy some of the flaws of other models while taking more seriously psychological constraints on input and processing. BEATS' access to input is limited to a single, left-to-right pass through the score, with limited 'memory.' Input to the model is a sequence of durations, without any of the other information that would be available to a listener (e.g. melody, harmony, accent) or to a reader of the score (e.g. bar lines or time signatures). The output is a metric hierarchy, represented in Lerdahl & Jackendoff's dot notation. BEATS can successfully handle scores with initial upbeats and/or syncopation.

BEATS was developed as a component of a larger, ongoing effort to implement GTTM as a computer simulation (Scarborough, Jones & Miller, 1989), but as a stand-alone model of meter extraction it is roughly comparable in scope to the models described above. BEATS draws most heavily on the production rules developed by Longuet-Higgins & Lee (1982), using them to generate candidate levels in the metric hierarchy. Their set of production rules has been expanded to provide three things: a way to generate the entire hierarchy; criteria for excluding levels generated

by the rules but not acceptable in the context of GTTM; and a means of generating those levels not generated by Longuet-Higgins & Lee's rules but required by the well-formedness rules of GTTM. A complete description of BEATS' production rules is found in Appendix A.

There is a much looser connection between BEATS and Povel's (1984) grid theory, in that both treat meter extraction as a matter of assessing the fit between a metric level, or grid, and a sequence of durations, but here all similarities end. Different means of assessing fit are used in the two approaches, and Povel is not concerned with the psychologically important question of how candidate grids are nominated. Whereas Povel's program selects, out of a large set, a single best-fitting grid, BEATS generates a family of grids representing a metric hierarchy that satisfies the GTTM metric well-formedness rules. This must be done as the program "listens" to the score rather than by processing the entire score at once, as Povel's algorithm does.

Basic Operations

In overview, BEATS has three types of processes: 1) Bottom-up processes that note time intervals between successive note onsets as they occur; 2) Top-down processes that take these intervals and combinations thereof and use them to predict the time of future events, with different

intervals leading to different predictions; and 3) Processes that evaluate the various predicted metric units or "grids" for consistency with a well-structured hierarchy as specified by GTTM.

Figure 0.1 shows BEATS' analysis of the beginning of Mozart's Symphony no. 40 (first violin part). It is comparable to GTTM's analysis (Lerdahl & Jackendoff, 1983, p. 23), with two differences. First, for reasons to be described, the analysis does not include the larger metric levels. Second, none of the metric levels in BEATS' analysis begins on the first note of the symphony, because a listener must hear a few notes before any metric structure emerges. Once BEATS has identified metric levels, it generates expectations at each level, like a listener who has "got the beat." Getting the beat, however, takes some time. Once this is done, BEATS generates a metric structure that conforms to the GTTM well-formedness rules: there is a dot at every note onset (rule 1); if there is a dot at a given level there is also a dot at the next lower (shorter duration) level (rule 2); dots at each level group dots at the immediately lower level either by twos or by threes (rule 3); temporal intervals between dots at any level are uniform (rule 4). Note also that strong beats (those with dots at higher levels) coincide with onsets of longer notes. This agrees

with Mozart's bar lines and with preference rules 4 (stressed notes should coincide with strong beats) and 5 (whenever possible, strong beats should coincide with, e.g., longer notes).

A central feature of BEATS is the time frame, a set of three equally-spaced times: T1, T2 and T3. When analysis begins, BEATS anchors the time frame at the onset of the first note by putting T1 there. When it reaches the beginning of the second note, it places T2 there. BEATS hypothesizes that the interval defined by T1 and T2 (the frame interval) is a significant metric unit in the music. This hypothesis predicts that another onset will occur one frame interval further on, so BEATS projects T3 one frame interval into the future, i.e. at T3 = T2 + frame interval. The hypothesis is made explicit in BEATS in a metronome, a process which 'ticks' at regular intervals to generate a level in the metric hierarchy. After the first note the frame interval is an ♪, so a metronome with a period of an ♪ is made. When this metronome ticks, it places a dot in the growing metric hierarchy and then sets itself to tick again one period later.

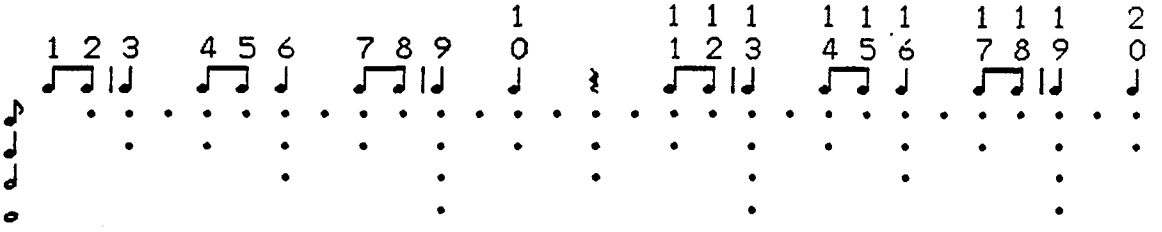


Figure 0.1 BEATS' analysis of Mozart's Symphony no. 40 (beginning).

BEATS is concerned with both musical events, i.e. note onsets, and metric events, i.e. metronome ticks, which predict note onsets. Accordingly, when BEATS has completed its processing at a given point, it moves on to the nearest event, whether note onset or metronome tick. In the present case, the distance to the sole metronome's next tick is the same as the distance to the next note onset (note 3), so BEATS moves to this location.

Generating Larger Levels

BEATS finds T3 at the new location, supporting the hypothesis that the ♪ is a metric unit. On the strength of this, BEATS hypothesizes a higher-level temporal grouping at double the metric unit just confirmed. There are two reasons for doubling rather than tripling a metric unit to generate the next higher level. First, ratios of two are more common. Second, two is the most common grouping in the subjective organization of identical, isochronous tones (Fraisse, 1982), and in spontaneous tapping (Fraisse, 1947-1948). The **Double** rule holds T1 fixed and moves T2 to where T3 is, thus doubling the frame interval, and projects

T3 to a new location one frame interval beyond T2. The new frame intterval generates a new metronome which ticks at ♩ intervals. There are now two metronomes, with different periods. BEATS examines each in turn to see if it is due to tick at the current location. In the example, both metronomes are due to tick at the current location (the third note onset), so now the metric hierarchy contains two dots at the ♪ level and one at the ♩ level.

The distance to the next note onset (note 4) is a ♩, while the next tick, which will be that of the ♪ metronome, is only an ♪ away. Accordingly, BEATS' next location is halfway between the onsets of notes 3 and 4. BEATS examines each metronome and finds that only the ♪ metronome is set to tick here.

Recognizing Upbeats

T3 occurs at the onset of the fourth note, so we might expect BEATS again to Double the frame interval and create a new metronome. However, something more important has happened that triggers a new rule. At the onset of note 4, BEATS finally knows the duration of note 3 (♩), and it recognizes that this note is longer than any it has heard before. Since longer notes usually initiate higher-level metric groupings (Povel & Essens, 1985), and since it is still early in the piece, BEATS retrospectively interprets the first two notes as upbeats to the third note. (The

scope of this retrospection, a parameter of BEATS, is five notes (Longuet-Higgins & Lee, 1982)). Accordingly, the **Upbeat** rule shifts the time frame forward, anchoring it so that T1 is at the onset of note 3, and T2 and T3 are one and two frame intervals (♩), respectively, from T1. The frame interval is not changed, and no new metronome is created. Note that the doubling that seemed warranted at this location is no longer possible, since T3 has been projected to a point we have not yet reached. BEATS now examines the metronomes and finds that both are due to tick.

The onset of note 5 corresponds to the tick of the ♪ metronome; the ♩ metronome does not tick here. At the onset of note 6, we have reached T3. Now the conditions of the doubling mentioned above are met. BEATS therefore Doubles the ♩ frame interval to a 𝅗𝅥, and makes a third metronome, with a 𝅗𝅥 period. Because of the precedence of the Upbeat rule over the Double rule, every second tick of this new metronome will coincide with a bar line in the score. At note 9 BEATS reaches T3 once again, where another doubling yields a frame interval of a whole note and a fourth metronome.

BEATS does not allow enlargement of the frame interval beyond a temporal limit which in practice is about the duration of a typical measure (e.g. a whole note).

Consequently, the analysis does not include GTTM's two-measure, four-measure and higher metric levels. The point of this limitation is that higher levels are: a) less perceptually salient (they are far above the tactus); and b) are better understood as defining phrasal boundaries than metric units (Longuet-Higgins & Lee, 1982). Indeed, at high levels GTTM allows for metric discontinuities and for violation of a well-formedness rule that normally requires dots at a given level to be equally spaced in time. Such intuitions have not yet been incorporated in BEATS.

Finishing the Analysis

Within the limits mentioned above, the four metronomes BEATS has created generate the metric hierarchy for the entire piece. This is accomplished by a production rule, Slide, which slides the time frame forward one frame interval whenever BEATS reaches T3 and no other rule applies. The Slide rule has the effect of moving the time frame through the piece in such a way that the present, i.e. the current location of BEATS, is always within the frame. The reason for doing this, rather than merely extending the existing metric levels through the whole piece, is that by continuing to predict future onsets on the basis of the largest experienced metric unit, the listener (and BEATS) may be able to detect metric changes

by the failure of the time frame's prediction. When a metric change occurs, the time frame may be reset at the point of change. This intuition has not yet been implemented in BEATS, but ultimately the problem of recognizing meter change will be an important test of any model of meter extraction.

Revising the Time Frame

The analysis of Mozart's Piano Sonata K.331, shown in Figure 0.2, illustrates another important operation. When BEATS gets to note 2 it puts T2 there, making the frame interval a 𝅘𝅥𝅯, and makes a 𝅘𝅥𝅮 metronome. At the onset of note 3, BEATS creates a ♪ metronome, reflecting the duration of the second note, but does not change the time frame. Before reaching T3, which is at the onset of note 4, BEATS realizes that note 3 is longer than the note beginning at T2, i.e. note 2. On the assumption that longer notes should initiate higher level metric units (Vos, 1977; Povel & Okkerman, 1981), the Stretch rule enlarges the frame interval to a ♩ by moving T2 to the onset of note 3. The time frame remains anchored to the onset of note 1. Unlike the Double rule, Stretch creates a new frame interval that is incompatible with the old one. Therefore the metronome corresponding to the frame interval that was Stretched (𝅘𝅥𝅯) is eliminated by the Remove rule. In addition, an ♪ level is created to fill the gap between

the ♩ and ♪ levels; this is described in the next section. Intuitively, the purpose of the Stretch procedure is to handle dotted notes. A dotted note is usually followed by a complementary note which, added to the dotted note, yields a duration that fits the metric hierarchy at a higher level than either of the notes alone.

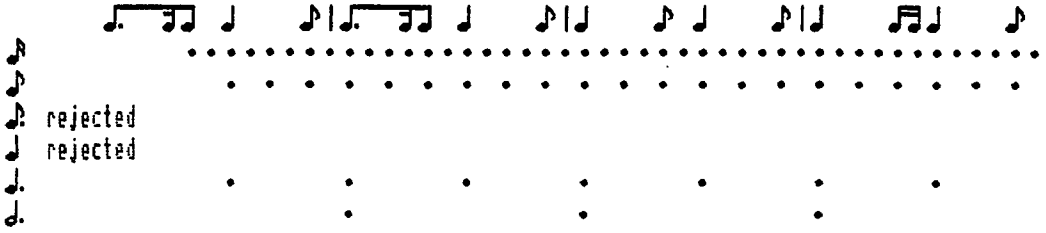


Figure 0.2 BEATS' analysis of Mozart's Sonata K. 331 (beginning).

BEATS has so far created 𝅘𝅥𝅯, 𝅘𝅥𝅰, 𝅘𝅥𝅱 and ♩ metronomes, and it has Removed the 𝅘𝅥𝅱 metronome. Halfway through note 4, BEATS reaches T3, disconfirming the time frame's prediction of a note onset. This suggests that the frame interval, although it has been Stretched, is still incorrect and needs to be enlarged again. Accordingly, Stretch moves T2 to the onset of note 4, yielding a frame interval of a ♩., and eliminates the metronome (♩) corresponding to the old frame interval.

Completing the Hierarchy

In the example above, BEATS created two metronomes (𝅘𝅥𝅯 and 𝅘𝅥𝅰) not because of a change in the time frame but because notes of those durations occurred in the piece. BEATS determines the onset interval between each pair of

adjacent notes, and creates a metronome corresponding to that interval if it is consistent with the current frame interval. Consistent here means that the onset interval in question is an integral multiple or divisor of the frame interval. For example, consider the beginning of the last movement of Mozart's Symphony no. 41 ("Jupiter"), shown in Figure 0.3. At the outset BEATS establishes a frame interval of a 𝅗𝅥, and only later hears the first of the ♩s. Longuet-Higgins & Lee's rules do not generate smaller metric levels, but it is assumed the music itself may directly dictate levels in such cases. Since the first ♩ is consistent with the frame interval, BEATS' Induce rule creates a ♩ metronome. A few notes later BEATS hears a ♩.; this, too, is smaller than the frame interval, but it is not consistent, so no metronome is made.

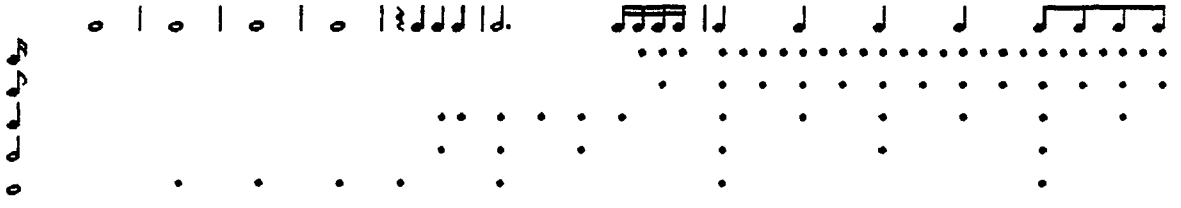


Figure 0.3 BEATS' analysis of Mozart's Symphony no. 41 ("Jupiter"), last movement.

The "Jupiter" example also illustrates a third way that metric levels are created. For the first four bars the metric hierarchy consists of a single level. When the ♩ metronome is created, its period is consistent with the current frame interval, but the metric hierarchy now

violates a well-formedness rule which limits the ratio between adjacent levels to three or less. The solution is the Interpolate procedure, which creates a new level which satisfies this rule with respect to both the ♩ level and the 𝅝 level, i.e. a 𝅗𝅥 level. After the ♩., we hear a ♪, and it is necessary to induce a ♪ metronome and to Interpolate an ♪ metronome.

(Appendix A provides a detailed description of BEATS' rules).

Conclusion

BEATS yields a psychologically and musically plausible metric analysis of music written in a strict meter (e.g. Bach, Mozart, American folk music, but not Stravinsky, Carter, Bulgarian folk music). Within this category there are many scores that it cannot correctly analyze. This is a limitation of duration-only analysis: duration is not the sole carrier of information about metric structure. The extent to which this approach succeeds reflects a degree of redundancy between rhythmic, melodic and harmonic dimensions in most music.

III. Testing the BEATS Model

To evaluate BEATS as a model of human perception, two kinds of questions can be asked. First, do listeners in fact extract a hierarchic representation of meter? Do they extract the same hierarchic representation that BEATS develops? Such questions are addressed in Experiments 1, 2 and 3. Second, if BEATS' metric hierarchies are in general a fair model of listeners' metric representations, are the specific rules used by BEATS useful in predicting listeners' perceptions? When BEATS makes an error do listeners make the same kind of error? These questions are addressed in Experiment 4.

All experiments used musically trained subjects. Lerdahl and Jackendoff (1983) describe their theory as reflecting the musical intuitions of a listener familiar with classical Western tonal music. To the extent that BEATS depends on this theory, it presumably can be said to reflect the metric intuitions of the same listeners. On the other hand the same cannot necessarily be said of the specific processes employed in BEATS, as Longuet-Higgins and Lee (1982) do not specify what kind of listener they are modelling. The most important reason for using musically trained subjects in these experiments was a practical one. In previous experiments naive subjects have in many cases not understood the nature of the task or the

aspect of the stimulus they were asked to judge. While appropriate instruction and practice could presumably overcome this problem, it seemed reasonable to begin with subjects to whom the tasks would make the most sense. Evaluating BEATS with respect to degree of musical training will have to await development of reliable methods for employing musically naive subjects.

1. The Metric Probe Technique

Experiments 1, 2 and 3 used the same set of stimuli and the same design, described below, in combination with three different tasks.

Rationale

In a study of the detectability of temporal displacements in a rhythmic context, Schulze (1978) used the following kinds of stimulus:

A. ¦ τ ¦ τ ¦ $\tau+\delta$ ¦ $\tau+\delta$ ¦ $\tau+\delta$ ¦ $\tau+\delta$ ¦

B. ¦ τ ¦ τ ¦ τ ¦ τ ¦ $\tau+\delta$ ¦ $\tau+\delta$ ¦

Brief pulses (¦) were separated by the interval τ or τ plus a small displacement, δ. There were either three (stimulus A) or five (stimulus B) pulses before the displacement was introduced. Subjects had to decide whether the sequence was regular (i.e. $\delta = 0$) or not. Of interest here is that displacements were more detectable when there were five pulses (stimulus B) before the displacement than when there

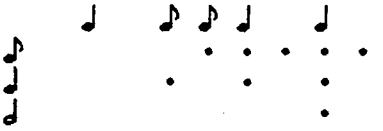
were three. This suggests that the initial pulses serve to induce a temporal expectation regarding the placement of future pulses. Each successive pulse increases the precision of the estimate of the inter-pulse interval (Schulze, 1989), with the result that deviations from this interval are more apparent in the five-pulse than in the three-pulse condition.

Bharucha & Pryor (1986) presented subjects with pairs of brief rhythms in a same-different experiment. In each pair, one rhythm was metrically exact and the other had one note displaced in such a way that it was not predicted by the existing metric structure. The inter-stimulus interval (isi) separating the two rhythms was a multiple of the basic metric interval defined by the first rhythm, as shown below (bars represent the rhythmic patterns, dots the basic meter):

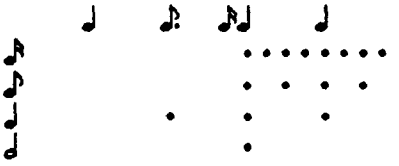
```
¦   ¦ ¦ ¦   ¦       ¦   ¦  ¦¦   ¦
• • • • • • • • • • • • • • • • • •
{---exact---}  isi  {-distorted-}
```

Discrimination was better when the exact rhythm preceded the distorted rhythm (as above) than when the distorted preceded the exact rhythm. The BEATS explanation of this phenomenon can be seen by representing the intervals above in musical notation.

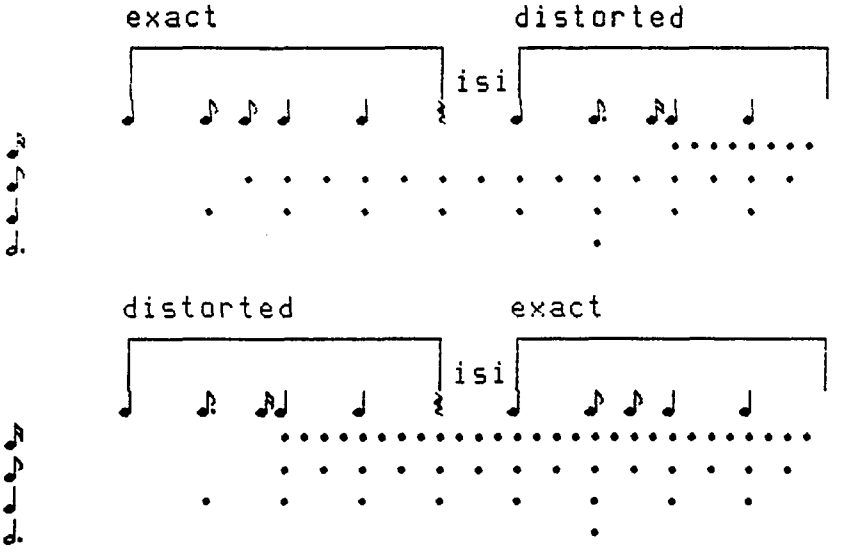
Exact:



Distorted:



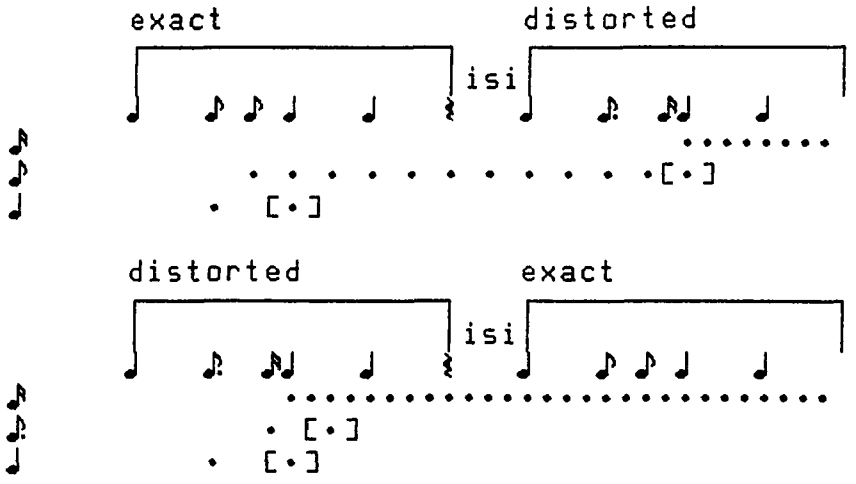
BEATS' analysis is shown beneath each rhythm. In the exact version, BEATS induces only ♪, ♩ and 𝅗𝅥 levels, while for the distorted version it induces 𝅘𝅥𝅯, ♪, ♩ and 𝅗𝅥 levels. In the latter case there is no ♪ in the rhythm, but a ♪ level is Interpolated between the 𝅘𝅥𝅯 and ♩ levels to complete the metric hierarchy. The asymmetry in discrimination can now be predicted by attaching the appropriate metric hierarchy to each stimulus order:



In each illustration the metric hierarchy is that generated by BEATS in response to the first stimulus; the second stimulus is shown superimposed on this metric structure at

the distance determined by the inter-stimulus interval (isi). In the first case the displaced note (♪) is not predicted by any metric level generated by the first (exact) stimulus, whereas in the second case the metric hierarchy induced by the first (distorted) stimulus predicts every note of the second rhythm. The subject's strategy might be to generate a hierarchy for the first rhythm and then evaluate its "fit" with the second rhythm. If the subject's hierarchies are like those generated by BEATS, such a strategy will clearly yield the asymmetry found by Bharucha and Pryor.

The above studies provide strong evidence for a fundamental principle of BEATS, namely that experienced periodicities generate expectancies on the listener's part of future note onsets. As shown above, the Bharucha and Pryor study is also consistent with the idea that these expectancies are hierarchically arranged. However, this is not the only possible interpretation of Bharucha's and Pryor's results. A simplified version of BEATS might generate the metric representations shown below.



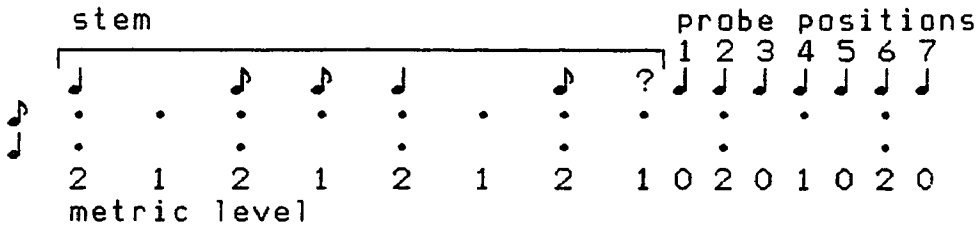
This model represents meter as a single expectancy about when the next note onset will occur. Whenever a new, smaller periodicity is encountered, the meter is changed. (The bracketed dots above indicate where the next beat would have occurred at a given level if the meter had not been changed.) This is a reasonable strategy if we suppose that the listener's task is to predict note onsets and that misses are more costly than false alarms. The point of this example is not to propose an alternative to BEATS but rather to point out that an essential aspect of BEATS' metric hierarchies is that all metric levels that BEATS creates are present <u>simultaneously</u> in the metric hierarchy. While BEATS accounts nicely for Bharucha's and Pryor's results, these results can also be explained by the simpler model described above. Experiments 1, 2 and 3 were an attempt to find more direct evidence of BEATS' metric hierarchies.

The metric probe stimuli used in all three experiments were the same and were based on two assumptions that have been discussed, namely:

1. Each metric level is a prediction of the time of the next note onset.

2. The more predictions a note onset confirms, the more strongly accented that note will be to the listener.

The basic idea of the metric probe technique is that various aspects of a listener's perception of a particular note ought to be affected by the number of predictions (if any) that note confirms. This is illustrated below.



The stimulus consists of a <u>stem</u>, a short rhythm that induces a particular metric structure, as shown, plus a <u>probe</u>, a single note added to the stem. The duration of the last note of the stem (denoted by ? above) varies, and may have any of seven possible values (from ♪ to ♩. in ♪ increments). As shown above, the probe may occur in one of seven possible locations, corresponding to the duration of the variable note. While all seven probe positions are shown in the above figure, the probe occurs in only one position in a given stimulus. The probe is arbitrarily

notated as a ♩ but in fact its duration is undefined, there being no following note. At the bottom of the figure is shown the number of confirmed predictions, or metric level, at each location in the stem and at each possible probe position. The metric level of a location at which no note onset occurs is the metric level that a note would have if one occurred there. A metric level of zero reflects the fact that a note onset at that location would confirm no predictions.

If the listener does in fact induce the metric structure predicted by BEATS, then the probe's metric level will be as described above. Therefore, by somehow asking the listener about the probe's metric level we can test BEATS' prediction. Because asking the listener about metric levels is necessarily an indirect inquiry, three tasks were used in the hope of providing converging evidence in support of BEATS' prediction.

1. Rating. Probes with higher metric levels might be perceived as fitting better with the stem (Palmer & Krumhansl 1987a, b). If so, asking subjects to rate the 'goodness' of the probe should yield a set of ratings that reflect the metric levels of the seven probe positions.

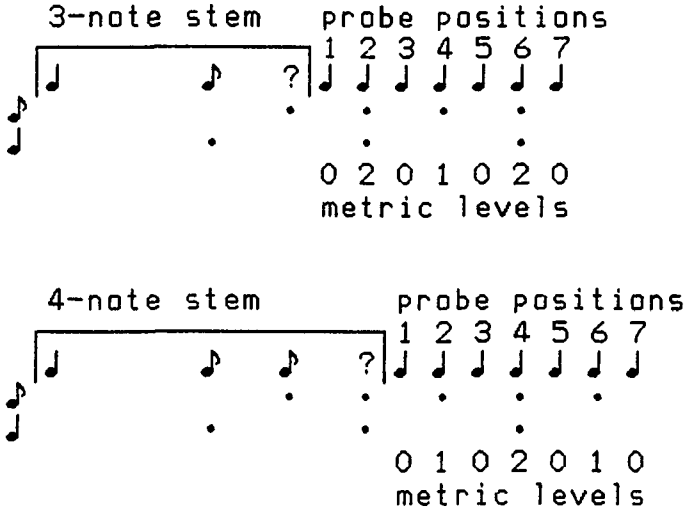
2. Identification. A general feature of perception is that that which is expected is more readily recognized and more accurately identified than that which is unexpected.

Probes with higher metric levels are more expected than probes with lower metric levels, and probes with metric levels of zero are completely unexpected. Accordingly, accuracy of identification of the interval preceding the probe (the variable note) should reflect the probe's metric level.

3. Reproduction. If greater expectancy leads to more accurate perception, then probes with higher metric levels should be more accurately reproduced than probes with lower metric levels. The relative accuracy of reproduction of the seven probe positions should therefore reflect their relative metric levels.

Metric Probe Stimuli

Each of the three metric probe experiments used the same set of stimuli, constructed as shown below. As described above, each stimulus consisted of a stem, ending in a note of variable duration, followed by a single probe note. The variable note ('?' below) had seven possible durations, ranging from a 𝅘𝅥𝅯 to a 𝅗𝅥. in 𝅘𝅥𝅯 increments. Two different stems were used, so as to vary the phase relations of the metric hierarchy and the probe positions, as shown below.



Because the length of the stem's last note (?) varies, the onset of the probe falls in one of the seven positions shown. Each such location is characterized by metric level, as determined by the metric hierarchy. An important feature of these stimuli for present purposes is the fact that metric level is not monotonic with probe position. Therefore any correlation between a dependent measure and metric level is not explainable in terms of increasing probe latency.

Stimuli replicated the conditions in which BEATS operates: all notes had the same articulation, intensity, pitch and timbre, with note duration the only cue to meter. A note consisted of a 40 millisecond pip followed by silence for the balance of the note's duration. Stimuli were produced using 440 Hz square waves, bandpass filtered at 400 and 1500 Hz to attenuate onset transients, and played over headphones at a comfortable listening level.

Design of the Metric Probe Experiments

All three experiments used the two stems illustrated above, with the same seven values for the variable note, yielding probes at three metric levels.

Each stem/variable note combination was presented once at each of three tempos: slow (75 ♩ beats per minute), moderate (106 bpm) and fast (150 bpm). These values were selected to span the range of common musical tempos; the middle value is such that the difference between slow and moderate and between moderate and fast is the same log ratio. In addition, a number of dummy trials were interpolated, played at randomly selected tempos in the same range. The object of this was to prevent learning of the six tempo/stem combinations, forcing the subject to attend to each stimulus. By increasing the variety of stimuli, the dummy trials reduced the likelihood that the subject remembered a response (in the rating and identification experiments) from one instance to the next of a given stimulus.

Each experimental session began with seven practice trials. Practice stimuli were constructed as described above, but the stems used were those shown below:

3-note practice stem: ♩ ♩ ?

4-note practice stem: ♩ ♩ ♩ ?

Following the practice stimuli there were 84 experimental

stimuli, constructed as described earlier, and 42 dummy trials. Of the 42 dummy trials, 28 combined the experimental stimuli (each stem x position combination was used twice) with randomly selected tempos. The remaining 14 dummy trials were constructed using random tempos in combination with the stems shown below:

3-note dummy stem: ♩. ♪ ?

4-note dummy stem: ♩. ♪ ♪ ?

The 84 experimental trials consisted of each stem x position x tempo combination presented twice, with the restriction that the entire stimulus set of 42 experimental stimuli (2 stems x 7 positions x 3 tempos) had to be presented before any member thereof was repeated.

The 126 stimuli (84 experimental plus 42 dummy) were arranged in six lists (three lists and their retrogrades). Each subject was run with each of the six lists, in an order determined by a Latin square, such that each list was used once in each of the six order positions.

2. Experiment 1: Rating the Metric Probes

Subjects

Six musically trained subjects participated in six sessions apiece. Each session lasted from 25 to 45 minutes. None of these subjects participated in either Experiment 2 or Experiment 3. These subjects ranged in age from late teens to early forties, with most in their

twenties; they included two brass players, two reed players, a pianist and a guitarist. Years of training ranged from 3 to 20, with a mean of 10.2 years. Four subjects had professional experience, and five had had training in rhythmic dictation.

Apparatus

Stimuli were presented, and responses collected, by a Turbo Pascal program, RPROBE, running on an AT-type computer. Machine language routines provided hardware-based millisecond timing resolution (Brysbaert, Bovens, D'Ydewalle & VanCalster 1989). The computer controlled a relay (Alpha Products A-Bus), which operated an electronic switch (Grason-Stadler 829E), gating the audio signal from a function generator (Clarke-Hess 743 and Hewlett-Packard 209A) with 10 millisecond rise and decay times. The signal was bandpass filtered (Krohn-Hite 3202) and played over small open-air type headphones (Realistic Nova-45). Subjects made rating responses by pressing a key on a small terminal connected to RPROBE via an ordinary serial connection.

Procedure

Subjects were run in a sound-attenuating booth inside a small, closed room. Communication between the subject and the computer was through the small terminal in the

subject's booth. The terminal's cursor was turned off to prevent its blinking from interfering with stimuli or responses.

Each session began with written instructions, followed by a series of practice trials. Subjects were told in advance that none of the practice stimuli would appear in the experiment. Subjects were not permitted to repeat an individual practice trial, but after the last practice trial they were given the option of repeating the practice session. The practice session could be repeated as often as the subject liked. Subjects were encouraged to repeat the practice session until they were comfortable with the apparatus and the procedure. Few subjects did in fact repeat the practice session, and none more than four times.

Before each practice trial, the subject was told either that the rhythm would be an example in which the last note fit well or that it would be an example in which the last note fit poorly. Subjects then heard the rhythm and were asked to rate how well the last note fit the rest of the rhythm, on a scale of 1 (poor fit) to 7 (good fit). No feedback was given, to avoid biasing the subject's use of the rating scale. The practice stimuli were chosen so as not to suggest hypotheses such as "if the last note is late (early) the fit is good (poor)," or "when the rhythm is long (short) the last note fits well (poorly)," and so

on. The relevant part of the instructions for the rating task are in Appendix B.

After the practice trials, subjects were asked if they had any questions, and then began the experimental trials. The trial cycle, described below, was designed to reduce subject fatigue by minimizing keystrokes and by pacing the experiment while allowing the subject to take a break when necessary. The trial cycle was as follows:

1. Inter-stimulus interval. The computer paused. During this interval the subject could temporarily stop the program (to rest briefly, reposition terminal, chair, etc.) by pressing the terminal's <ESC> key. This resulted in a message on the terminal instructing the subject to hit any key to resume. Alternatively, the subject could end the inter-stimulus interval immediately by pressing the space bar. Finally, the duration of the inter-stimulus interval could be toggled between long (two and a half seconds) and short (one second) by pressing the 'S' (slow) and 'F' (fast) keys during this interval.

2. Warning tone. The terminal beeped. The computer did not respond to any of the keys mentioned above after this.

3. Attention interval. Following the warning tone was a pause for a random period in the range 1000 - 1400 milliseconds. Any duration in this range is longer than the first interval of any of the stimuli. This, in

conjunction with the fact that the warning tone was considerably higher pitched than the stimulus tone, prevented the warning tone from becoming part of the stimulus.

4. Stimulus presentation.

5. Response prompt. A message on the terminal screen asked for a rating to be entered from the keyboard. Only the keys 1 - 7 were echoed to the terminal screen; pressing any other key elicited a beep from the terminal.

6. End of response. The subject pressed <ENTER> to terminate the trial. Before pressing <ENTER>, the subject could change his/her choice any number of times by pressing another key. The computer signalled the subject that the response period was over by clearing the terminal screen.

Results and Discussion

Because distributions of ratings tend to be skewed, medians were used in place of means in analyzing these data. Median ratings for each stem x level combination, pooled across positions, subjects and tempos, are shown in Table 1.1. Median ratings for each stem x probe position combination, similarly pooled, are shown in Table 1.2.

median rating		metric level 0	1	2	median
stem length:	3	3.3	4.3	5.8	4.2
	4	3.8	5.0	6.6	4.6

Table 1.1 Median ratings by metric level.

median rating stem length	probe position 1	2	3	4	5	6	7
3	3.0	6.6	2.9	4.3	3.9	4.8	4.2
	(0)	(2)	(0)	(1)	(0)	(2)	(0)
4	3.4	6.0	4.6	6.6	4.1	3.1	4.6
	(0)	(1)	(0)	(2)	(0)	(1)	(0)

(metric level in ( ))

Table 1.2 Median ratings by probe position.

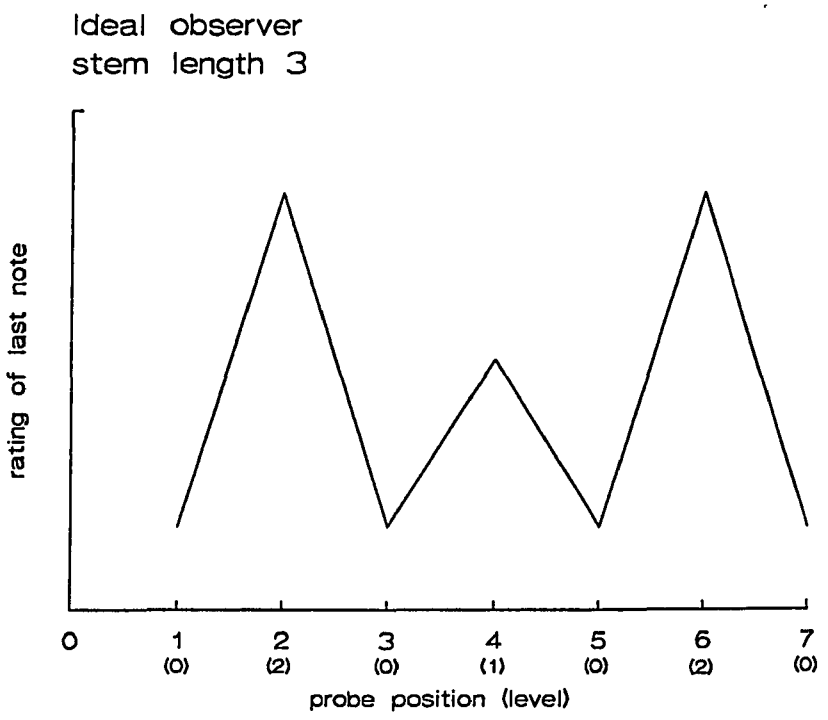
Because session and stimulus order are not independent, each factor was evaluated in a separate repeated-measures analysis of variance. No significant effects were found, and both factors were eliminated from further analysis. Median ratings were analyzed in a 2(stem) x 7(position) x 3(tempo) analysis of variance design with repeated measures on all factors; because probe position and metric level are confounded, a second analysis replaced the position factor with the level factor. The effect of probe position is significant ($F_{6,30} = 50.05$, $p < .01$) as is the effect of metric level ($F_{2,10} = 51.27$, $p < 0.01$). There is a weak interaction of probe position with tempo ($F_{12,60} = 4.82$, $p < .01$) but no corresponding interaction of level with tempo. The effect of stem length is not significant, but the higher ratings for rhythms with the longer stem (see Table 1.1) are as predicted by Schulze's (1978) results, discussed earlier. The fact that three positions (2, 4 & 6) have different metric levels in the two stems predicts an interaction between stem and position ($F_{6,30} = 4.80$, $p <$

.01), as shown in Table 1.2. These results seem to support the BEATS hypothesis, but further analysis shows that the picture is more complicated.

Figure 1.1 shows the responses of an ideal observer, as predicted by the BEATS hypothesis, for each stem. BEATS predicts that all probes at metric level 2 will be rated higher than all probes at level 1, which will in turn be rated higher than all probes at level 0. This much is borne out by the data, as shown in Figure 1.2 and in Table 1.1.

The BEATS hypothesis also predicts that all probes at a given level will receive more or less the same rating. This prediction is not entirely supported by the data (see Figure 1.2). Finally, BEATS does not predict that subjects will differ in their ratings, but clearly they do, as shown in Figure 1.3, which shows each subject's median ratings for the three-note stem, and in Figure 1.4, which shows the same for the four-note stem.

Before concluding that the data do not support the BEATS model, it is worth noting two systematic features of the data. First, considering the ideal observer's ratings depicted in Figure 1.1, the points at positions 2, 4 and 6 can be thought of as peaks in a function relating position and rating. Now considering the subjects' data in Figures 1.3 and 1.4, it is clear that subjects do not all have



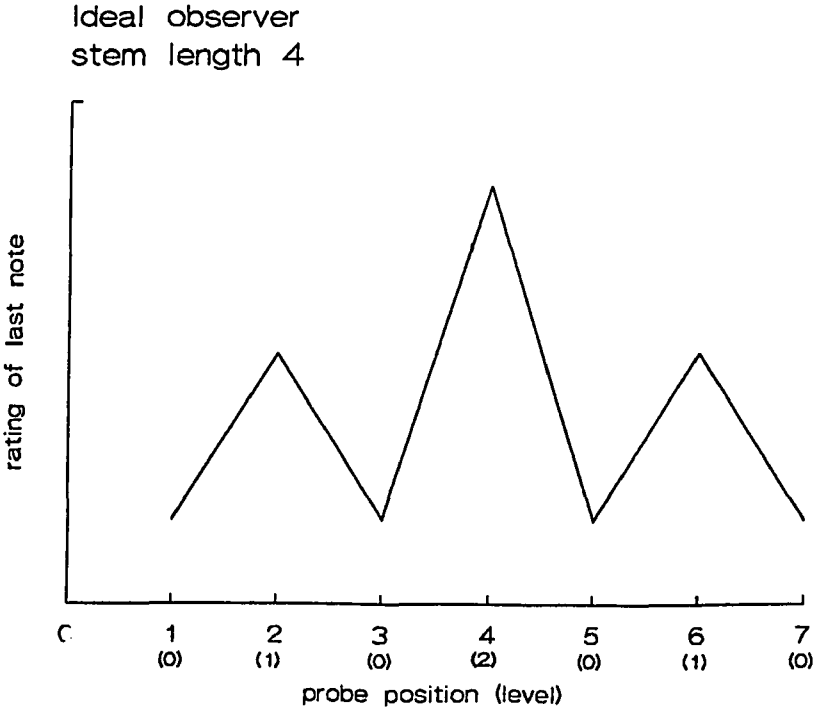
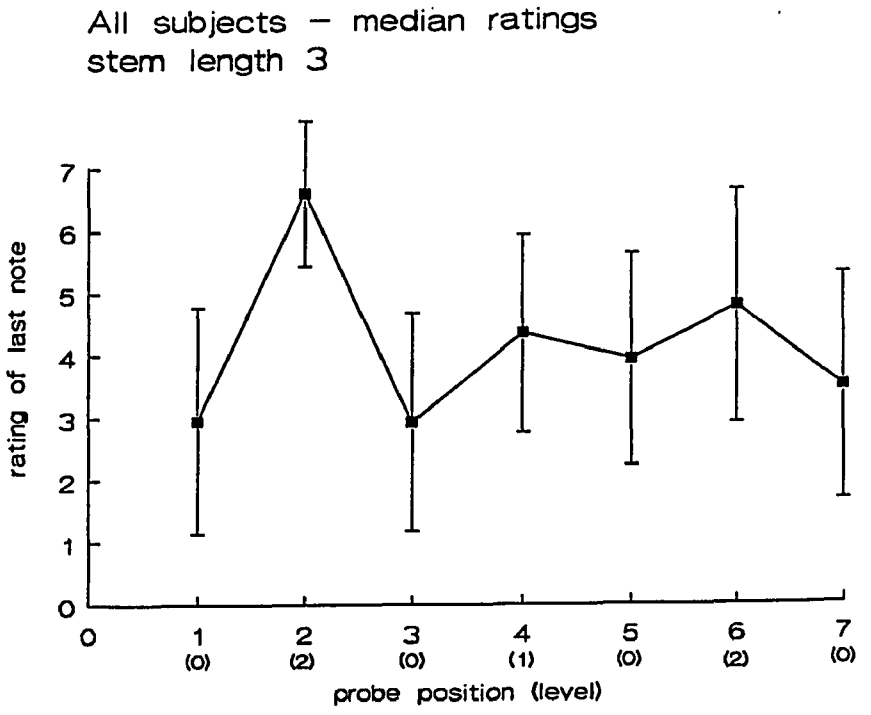


Fig. 1.1 Ratings by an ideal observer.



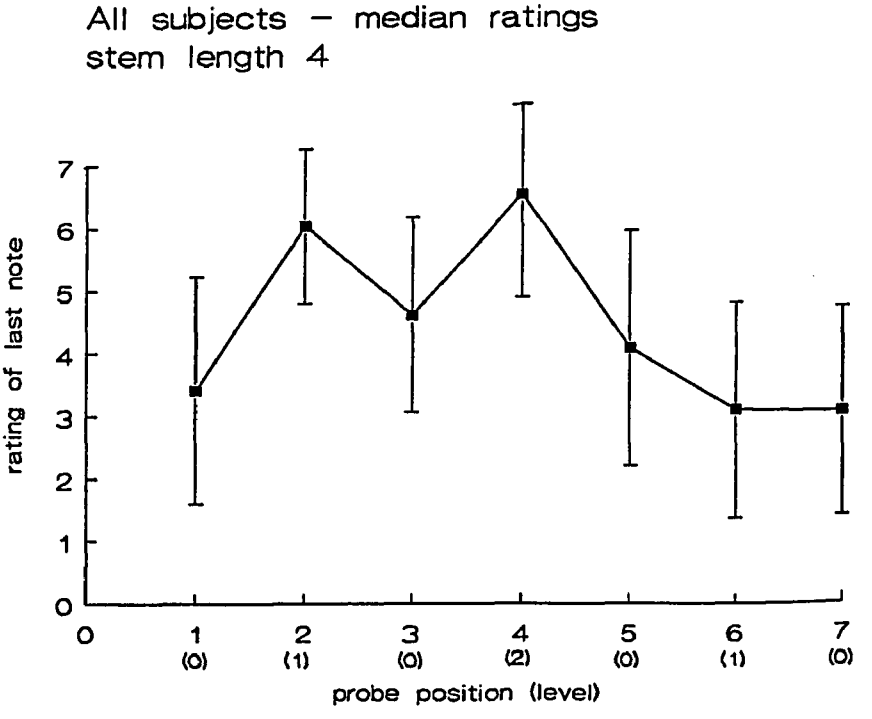


Fig. 1.2 Median ratings for all subjects. Error bars are standard deviations.

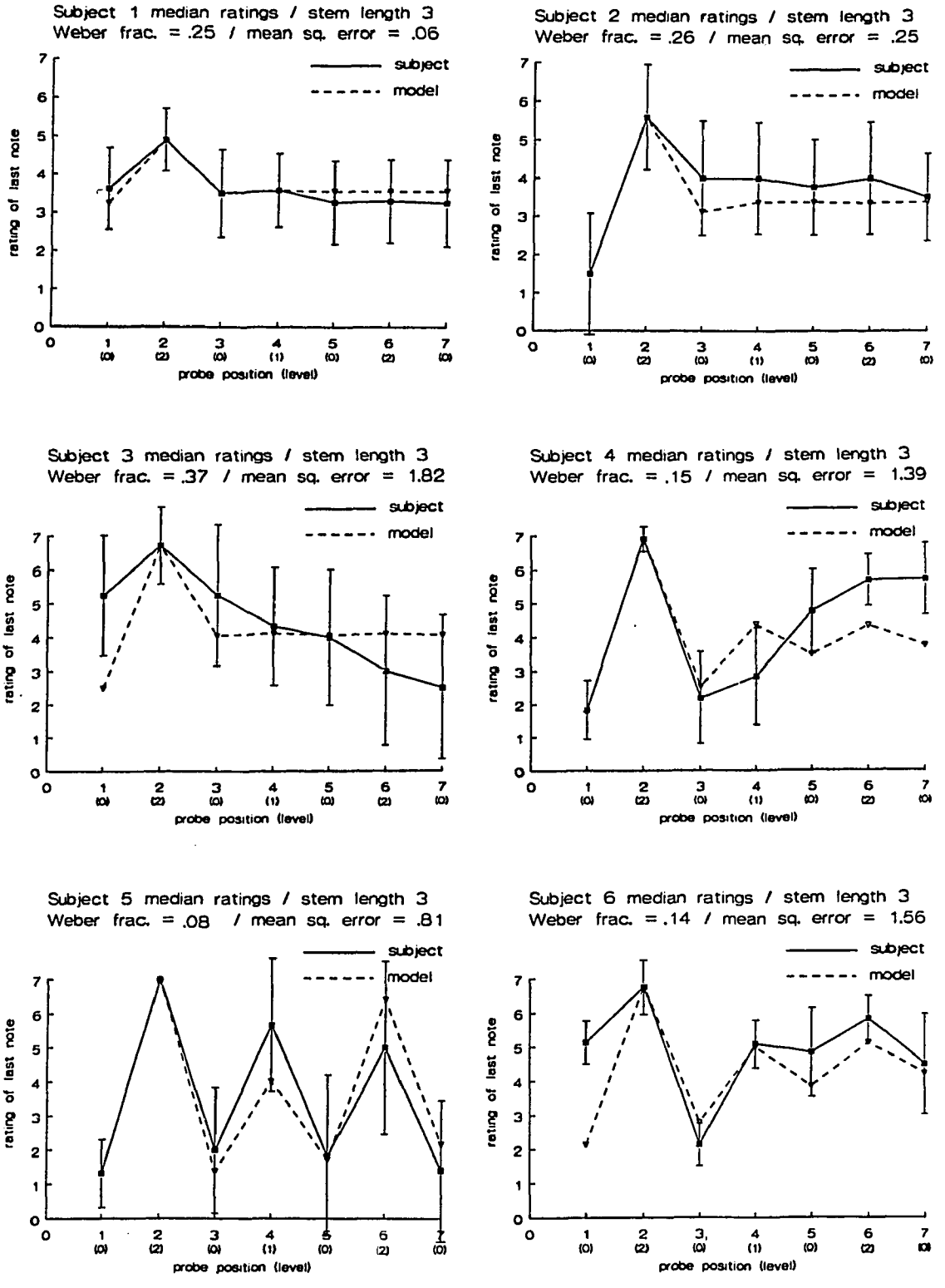


Fig. 1.3 Median ratings for each subject for three-note stem. Error bars are standard deviations; points without error bars have no variance. Weber fractions and mean squared errors describe the models fit to subjects' data.

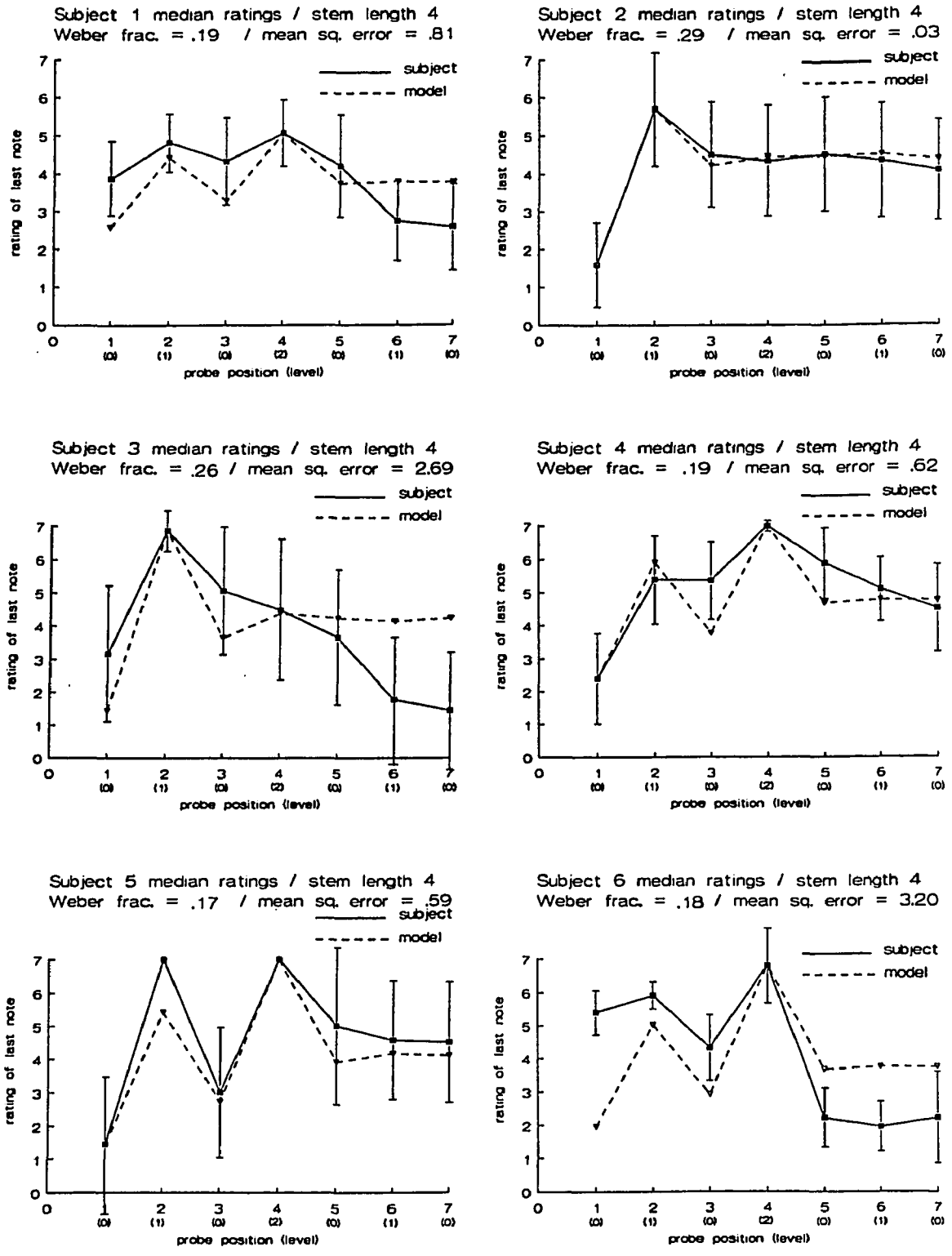


Fig. 1.4 Median ratings for each subject for four-note stem. Error bars are standard deviations; points without error bars have no variance. Weber fractions and mean squared errors describe the models fit to subjects' data.

peaks where peaks are predicted. However, the peaks that subjects do have are in predicted positions; no subject has a peak where one is not predicted. Second, subjects' ratings were more likely to reflect different metric levels in the early positions than in the later positions. Subject 1, for instance, distinguished between the metric levels of the first three positions for the three-note stem but not between any of the other positions. These two characteristics of the data can be accounted for by a psychophysical model of the subject's task. The model can account for much of the inter-subject variance and can reconcile these data with the idea of metric hierarchies as psychological representations.

The Weber's Law Model

The model divides the subject's task into three stages. The first stage covertly maps the temporal interval from the end of the stem to the onset of the probe onto a representation of the temporal dimension of the incoming stimulus. The second stage then maps this temporal dimension onto the metric hierarchy, thus identifying the metric level of the perceived temporal position. The third stage then maps the identified metric level onto the rating scale, by means of a mapping function that presumably incorporates subject bias.

The upper part of Figure 1.5 shows Stage 1 of the

Weber's Law model in more detail. Weber's law predicts that perception of the temporal position of the probe will become poorer as the interval preceding the probe grows (Lunney, 1974; Halpern & Darwin, 1982). As shown in Figure 1.5, this interval grows by a constant increment from one probe position to the next. Therefore the decline in temporal resolution is due to the increasing spread of the sensory distribution of the probe's position. Stage 1 maps the seven probe positions onto a continuous temporal dimension. The later the probe, the greater the range of the temporal dimension onto which it can be mapped. Clearly, differences in the discriminability of the probe position, both from position to position and from subject to subject, will have consequences for identifying the probe's metric level.

These consequences show up in Stage 2. As shown in the lower part of Figure 1.5, the continuous temporal dimension is mapped onto the discrete positions of the metric hierarchy produced by BEATS, represented by the spaces between dotted vertical lines. Each of these positions corresponds to a region of the temporal dimension. The sensory distributions of Stage 1 determine the probability that a probe will be perceived in a particular position on the temporal dimension. By calculating the area under each sensory distribution for each probe position (defined by a

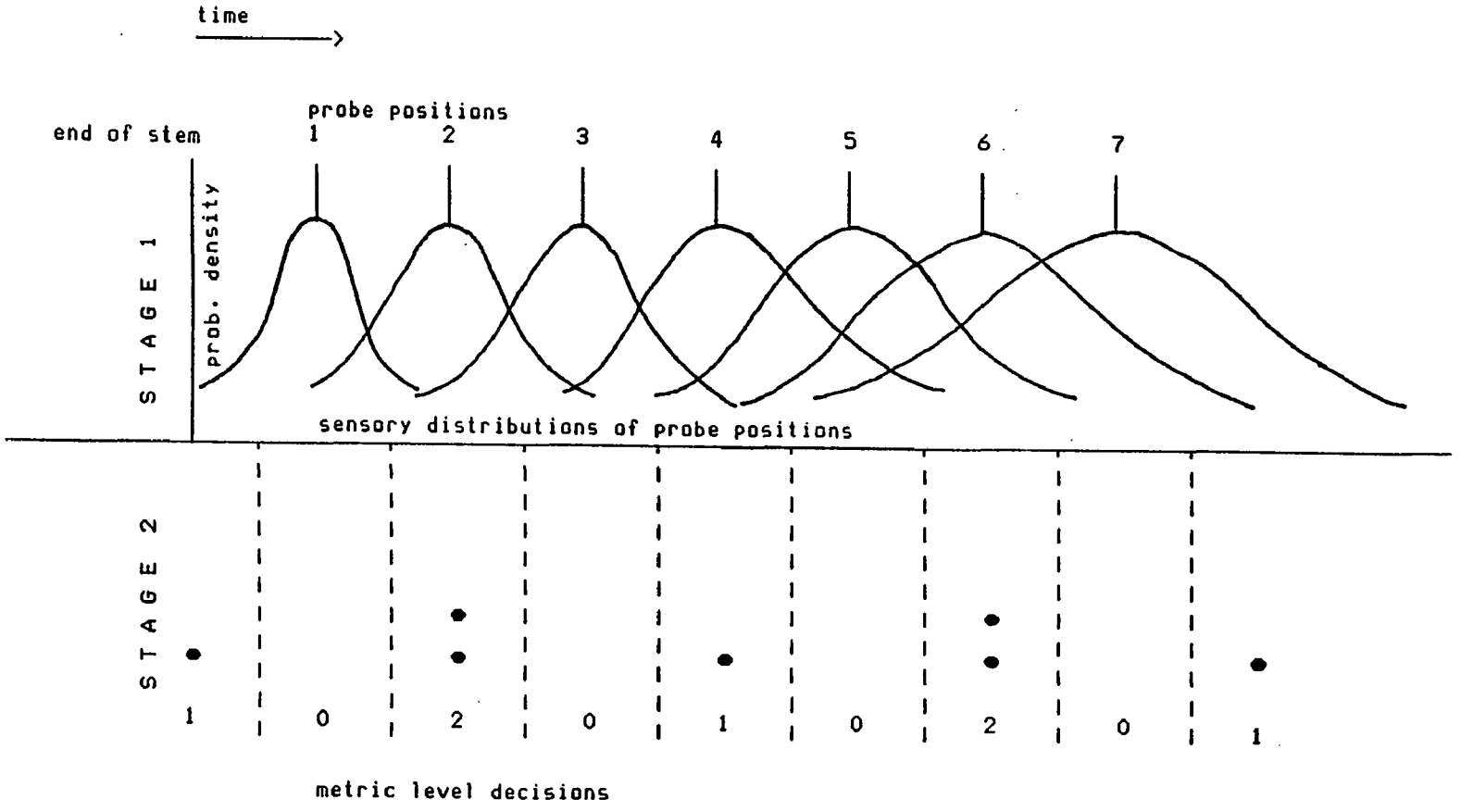


Fig. 1.5 Stages 1 and 2 of the Weber's Law model.

pair of dotted lines), it is possible to derive the probabilities of each probe position being perceived at each of the three metric levels. These three probabilities can then be combined in a single median metric level for each probe position. For example, if the sensory distribution for position 2 has a standard deviation of 0.5 positions, then the probability that the probe will be perceived to fall at metric level 2 is given by the area under the curve between the dotted lines defining position 2, which is .6826. The probability that the probe will be perceived to fall at metric level 0 is given by the two areas under the same curve between the pairs of dotted lines defining positions 1 and 3, which are .1574 each, or .3148. Likewise, the probability that the probe will be perceived to fall at metric level 1 is given by the two areas under the same curve between the pairs of dotted lines defining positions 0 and 4, which are approximately .0013 each, or .0026. This probability distribution is summarized below.

level	prob.	cum. prob.
0	.3148	.3148
1	.0026	.3174
2	.6826	1.0

Just as subjects' responses are characterized by a median rating, a median metric level can be used to characterize the above distribution. The median is used here because the distribution of metric-level identifications at a given

probe position tends to be non-normal and skewed, as shown above. By defining the median of such a distribution as the metric level value with a cumulative probability of 0.5, it can be easily calculated. In the example, this value is 1.75. That the median metric level is less than two reflects the fact that, according to the model, a probe in position 2 will usually be perceived in the temporal region corresponding to a position in the metric hierarchy at level 2, but on a fair number of trials it will be perceived in one of two regions corresponding to metric level 0. The model's median metric level also reflects the spread of the position 2 sensory distribution: the greater the spread, the lower the median.

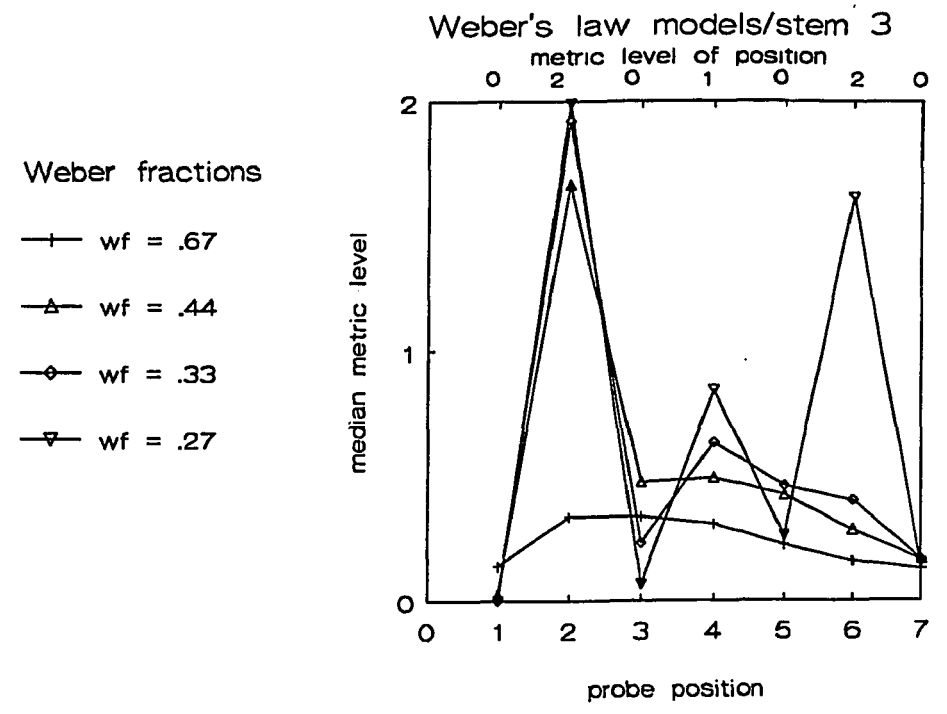
The Weber's law model has a single parameter, namely the Weber fraction. A listener with a small Weber fraction has smaller sensory distributions, and therefore better position discrimination, than a listener with a larger Weber fraction. The appeal of this model comes from the fact that changes in discrimination, as reflected in the Weber fraction, interact in a very interesting way with the metric hierarchy.

An ideal Weber's law model observer can be simulated for any Weber fraction. The fraction determines the relative dispersion of the seven sensory distributions. The median metric level of a probe in each position can

then be computed as described above. Figure 1.6 shows the rating patterns of a family of simulated ideal Weber's law model observers differing only in the Weber fraction parameter. What is remarkable about these patterns is that they give rise to the same observations made earlier about subjects' data, namely that prominent peaks occur only where they are predicted, and that judgments of fit are most likely to show differences for early positions.

By adjusting the Weber fraction parameter, the best fitting model can be found for each subject, thus allowing the Weber's law model to account for individual differences in temporal resolution as well as the systematic features of the data discussed above. Each Weber's law model described below was fit to the data by iteratively searching for the Weber fraction that gives rise to the set of median metric levels (computed as described above) that best fit (using a least squares criterion) the subject's median ratings.

In order to assess the fit between the Weber's law models and the subjects' data it is necessary to make some assumptions about the way metric level is mapped onto the rating scale in Stage 3 of the model. By assuming that the ideal Weber's Law model observer is unbiased and uses the entire rating scale, it is possible to map the model's 0 - 2 range of median metric levels directly onto the 1 - 7



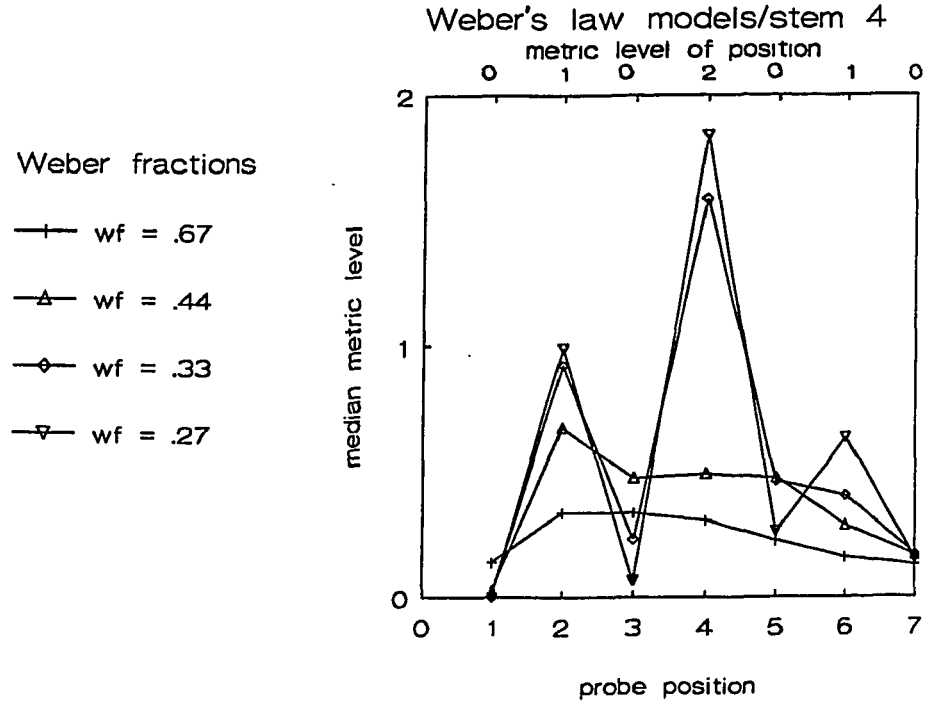
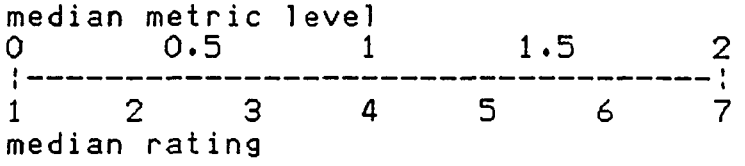


Fig. 1.6 Median stage 2 metric levels in four models.

range of the rating scale. Thus a (model's) median metric level of 0 corresponds to a (subject's) rating of 1, a level of 2 corresponds to a rating of 7, and intermediate median metric levels correspond to intermediate median ratings, as shown below.



Because they do not produce ratings, the Weber's law ideal observers do not model response biases or differences between subjects in using the rating scale. With this constraint, correlations between model output and subject data show that the model accounts for between 15% and 83% of the variance within a given subject's data; the mean is 53%. There is no relation between a subject's sensitivity, as reflected by the modeled Weber fraction, and the amount of variance explained by the model ($r = -0.08$). Thus to the extent that the model works, it works equally well for all subjects, which is as it should be given that the model purports to explain individual differences.

While for some subjects the model accounts nicely for the data, for others it does rather poorly, raising the problem of explaining the remaining variance. Certainly some of this variance could be explained if the model incorporated a third stage that reflected the response bias

of real, rather than ideal, observers. While it is not clear how to do this in a formal way, an idea of the effects of doing so can be had by scaling the model's predictions so that, in terms of the metric level-to-rating mapping described above, the model's lowest and highest metric levels correspond to the subject's lowest and highest median ratings and the model's mean metric level corresponds to the subject's mean rating. Models scaled in this way are plotted in Figures 1.3 and 1.4 together with subjects' data. The fit of both ideal and scaled models is shown in Table 1.3.

		----ideal----			----scaled---			
subject	stem	WF	MSE	r^2	WF	MSE	r^2	
1	3	.38	1.32	.67	.25	0.06	.82	
1	4	.19	2.68	.24	.19	0.81	.17	<-
2	3	.36	1.31	.77	.26	0.25	.92	
2	4	.20	3.05	.64	.29	0.03	.98	
3	3	.37	4.31	.28	.37	1.82	.19	<-
3	4	.19	4.07	.14	.26	2.69	.27	<-
4	3	.12	2.94	.59	.15	1.39	.63	
4	4	.17	5.99	.62	.19	0.62	.74	
5	3	.09	0.96	.83	.08	0.81	.84	
5	4	.14	4.10	.77	.17	0.59	.92	
6	3	.12	4.77	.56	.14	1.56	.48	<-
6	4	.17	4.62	.24	.18	3.20	.19	<-
mean:	3	.24	2.60	.62	.21	0.98	.65	
	4	.18	4.08	.44	.21	1.32	.54	

Table 1.3 Weber fraction (WF), mean squared error (MSE), and fit (r^2) of ideal and scaled Weber's Law models for the rating subjects.

Scaling the models' output improves their fit somewhat. The models fulfill, to some degree, the expectation that the Weber fraction be the same for the two stem lengths for a given subject: the correlations between Weber fractions

for the three-note stem and for the four-note stem are 0.87 for the ideal models and 0.76 for the scaled models. The five worst-fitting models, indicated by left arrows (<-) in Table 1.3, are also the only five models in which the positions of the model's lowest and highest ratings do not correspond to those of the subject. Such cases in particular make it clear that Stage 3 must do more than linear transformations if it is to explain differences among subjects.

3. Experiment 2: Identifying the Metric Probes

Subjects

Six musically trained subjects participated in six sessions apiece. One subject did not complete the experiment, explaining that she found it difficult to pay attention (her data confirm this), and a seventh subject was recruited as a replacement. Each session lasted from 25 to 50 minutes. None of these subjects participated in either Experiment 1 or Experiment 3. These subjects were in their twenties and thirties, except for one subject in her sixties; they included two keyboardists, a guitarist, a singer, a percussionist and a flutist. Years of training ranged from 12 to 20, with a mean of 14.75. Three subjects had professional experience, and the same three had had training in rhythmic dictation.

Procedure

Because of the pronounced reluctance of many potential subjects to make the journey to Brooklyn College (where all the above-described apparatus is housed), it was necessary to make the identification experiment portable. Accordingly, the stimuli were recorded, using the setup described above, and response sheets, one for each stimulus order, were made up. Subjects were run in various locations (including, as it turned out, Brooklyn College) that had the same characteristics as the Brooklyn College location, namely a) a quiet room, and b) no possibility of interruption by people or telephones. Subjects sat at a table, across from the experimenter, approximately three feet from the portable tape player used to present the stimuli.

Subjects were asked to identify the interval preceding the probe note. The duration of the first note of the rhythm (♩ or ♩.) was provided to the subject after the stimulus was presented. Each trial corresponded to a line on the response sheet. For example:

trial 27 first note = ♩ 𝅘𝅥𝅯 𝅘𝅥𝅯 𝅘𝅥𝅯 ♩ 𝅘𝅥𝅯𝅘𝅥𝅯 ♩. ♩.

The seven note values to the right are the response choices; these were the same on every trial. The subject kept the line for the current trial covered with a card until the stimulus had been played. At that point the

subject moved the card down one line, exposing the note value of the first note of the stimulus as well as the response choices. Subjects responded by circling the appropriate choice. Subjects were instructed to make their responses quickly, and were told a) that intervals corresponding to each of the note values in the response choices would in fact occur, and b) that they should not necessarily attempt to use all responses equally often.

Each session began with the same practice trials used in the rating and reproduction experiments, and, as in those experiments, subjects were allowed to repeat the practice session as many times as they wanted to. After the practice trials subjects were asked if they had any questions, and then began the experimental trials. The trial cycle, described below, was comparable to that of the rating and reproduction experiments in that it minimized actions on the subject's part and paced the experiment while allowing the subject to take a break when necessary. The trial cycle was as follows:

1. Inter-stimulus interval. There was a three second pause between rhythms on the tapes, during which the experimenter pressed the pause button on the tape player. Once the subject had made a response and indicated readiness for the next trial (or requested a break), the pause button was released.

2. Warning signal. The slight click caused by releasing the pause button, as well as the hiss of the tape, served to notify the subject that another rhythm was imminent.
3. Attention interval. Once the pause button was released, the remainder of the three second pause (on the tape) was in effect a random attention interval.
4. Stimulus presentation.
5. Response. The pause button was pressed. The subject slid the card down and made a response by circling the appropriate note symbol(s).

Subjects were given four short breaks in the course of each session when the experimenter interrupted the above cycle to give the subject the next response sheet.

Results and Discussion

An arcsin transform (Winer, 1971) was used to remove the correlation between cell means and cell variances of the proportions of correct responses. Preliminary analysis showed no effects of session or stimulus order, and these factors were eliminated from further analysis. Arcsin-transformed proportions of correct responses were analyzed in a 2(stem) x 7(position) x 3(tempo) analysis of variance design with repeated measures on all factors; a second analysis replaced the position factor with the level factor. Only the main effects of probe position ($F_{6,30}$ = 19.24, $p < .01$) and probe metric level ($F_{2,10}$ = 11.27, $p <$

.01) are significant. Mean proportions of correct responses for each stem x level combination, pooled across positions, subjects and tempos, are shown in Table 2.1. Mean proportions of correct responses for each stem x probe position combination, similarly pooled, are shown in Table 2.2.

mean p(C)		metric level 0	1	2	mean
stem length:	3	.39	.50	.73	.50
	4	.43	.69	.77	.55

Table 2.1 Mean proportions of correct responses by metric level.

mean p(C)	probe position 1	2	3	4	5	6	7
stem length 3	.89	.86	.31	.50	.16	.61	.18
	(0)	(2)	(0)	(1)	(0)	(2)	(0)
4	.85	.85	.40	.77	.26	.52	.21
	(0)	(1)	(0)	(2)	(0)	(1)	(0)

(metric level in ( ))

Table 2.2 Mean proportions of correct responses by probe position.

While there is no significant effect of stem length, the higher proportions of correct responses for the four-note stem are consistent with the higher ratings for that stem in the previous experiment. Figure 2.1 shows each subject's proportion of correct responses for each position for each stem, pooled across tempos, and Figure 2.2 shows the same data pooled across subjects. The stimulus/response confusion matrix, pooled across subjects, stems and tempos, is shown in Table 2.3.

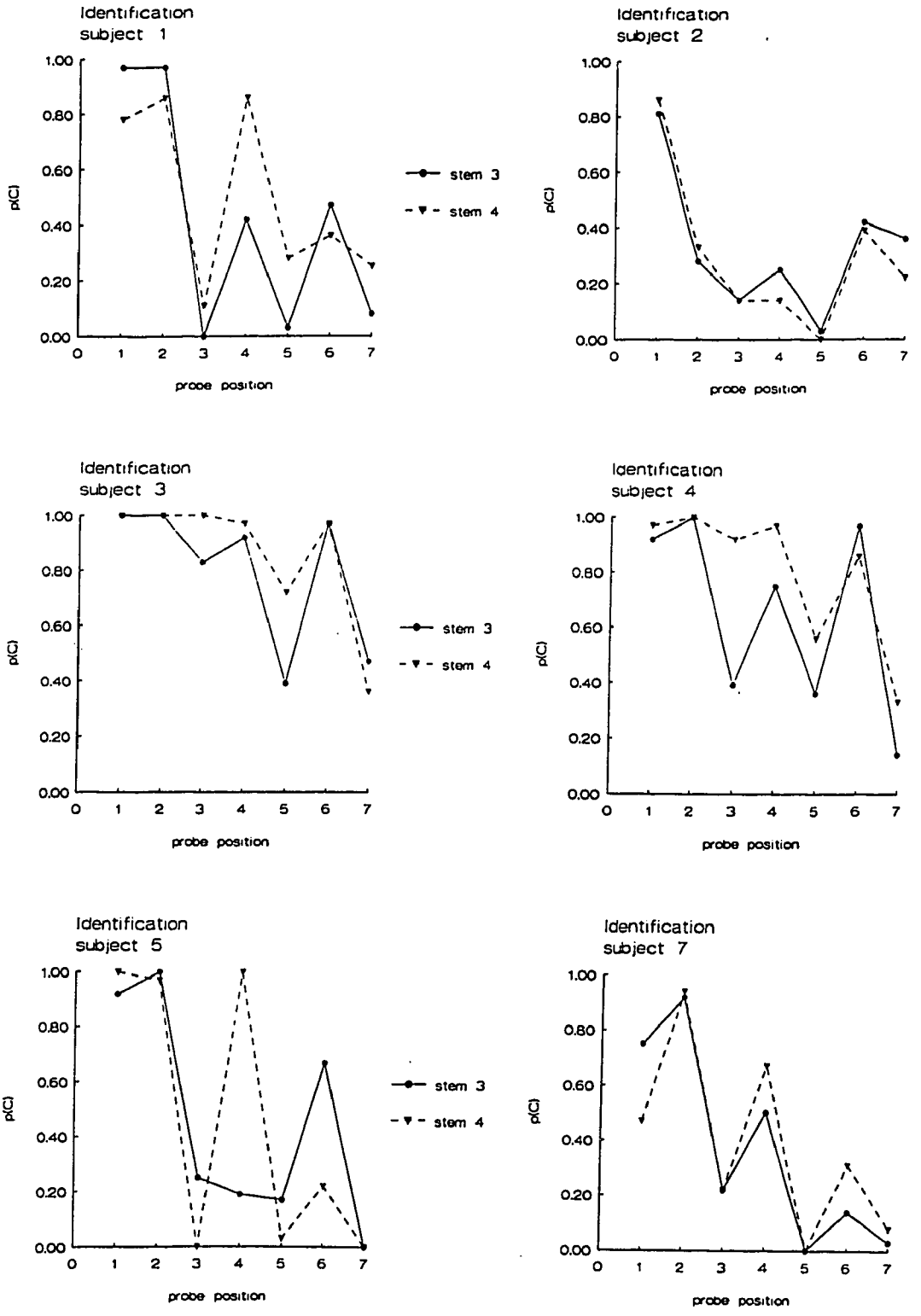


Fig. 2.1 Mean proportion correct for each subject in the identification task.

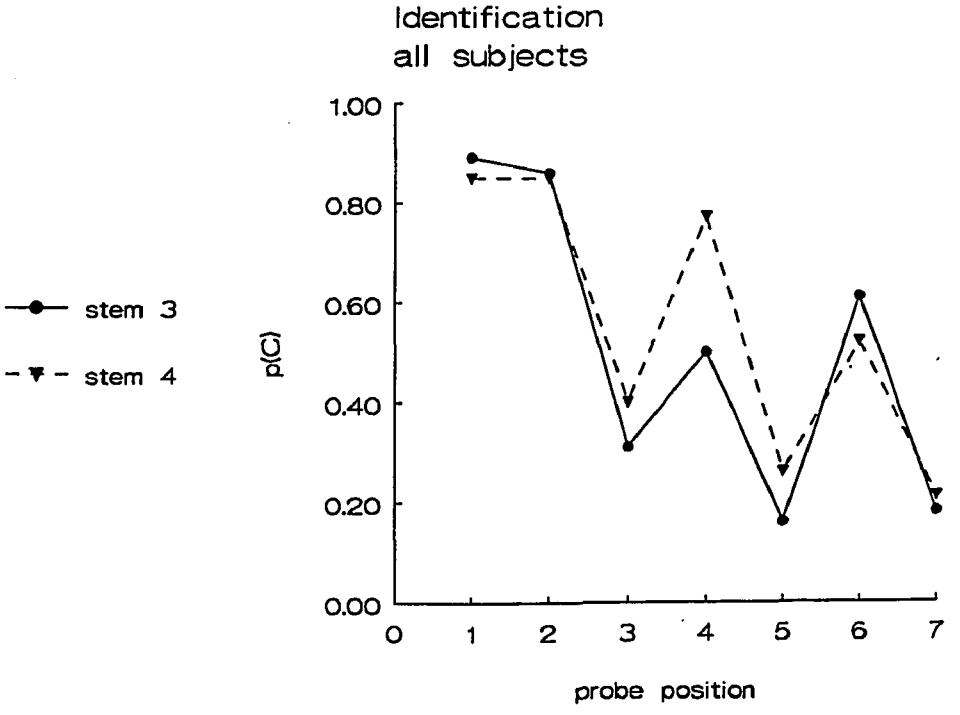
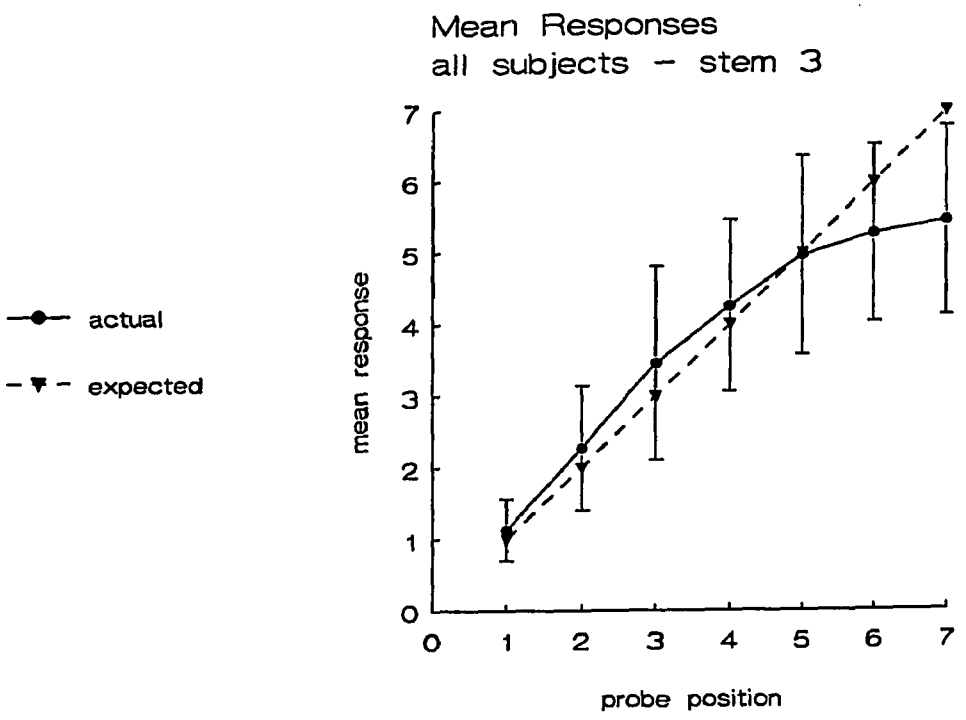


Fig. 2.2 Mean proportion correct for all subjects in the identification experiment.

frequency	probe position 1	2	3	4	5	6	7	total	rank
response 1	376	10	5	3	5	2	1	402	4
response 2	49	370	78	25	17	11	14	564	3
response 3	5	15	152	40	26	25	25	288	5
response 4	2	28	120	278	129	80	52	686	2
response 5	0	2	28	15	92	45	46	228	6
response 6	0	5	42	69	141	243	210	710	1
response 7	0	2	7	5	22	26	84	146	7
	432	432	432	432	432	432	432		

Table 2.3 Stimulus/response confusion matrix.

There are several general features of these data. First, performance generally declines with increasing delay of the probe note. This is consistent with the general predictions of the Weber's law model, in which confusion among neighboring positions is greater the later the probe position. This can be seen in the confusion matrix and in Figure 2.3, in which the leveling-off of the curve indicates smaller differences in mean responses for later probe positions, indicating that later probe positions are less well discriminated from one another than earlier positions. This is also reflected in the growth of error bars through the first three positions. It is possible that error reaches a maximum at position 3, but the fact that subjects rarely used responses 6 and 7 limits the size of the error bars for the later positions. The confusion matrix, on the other hand, shows that error, in terms of deviations from correct responses, grows over the whole range of probe positions, as predicted by Weber's law.



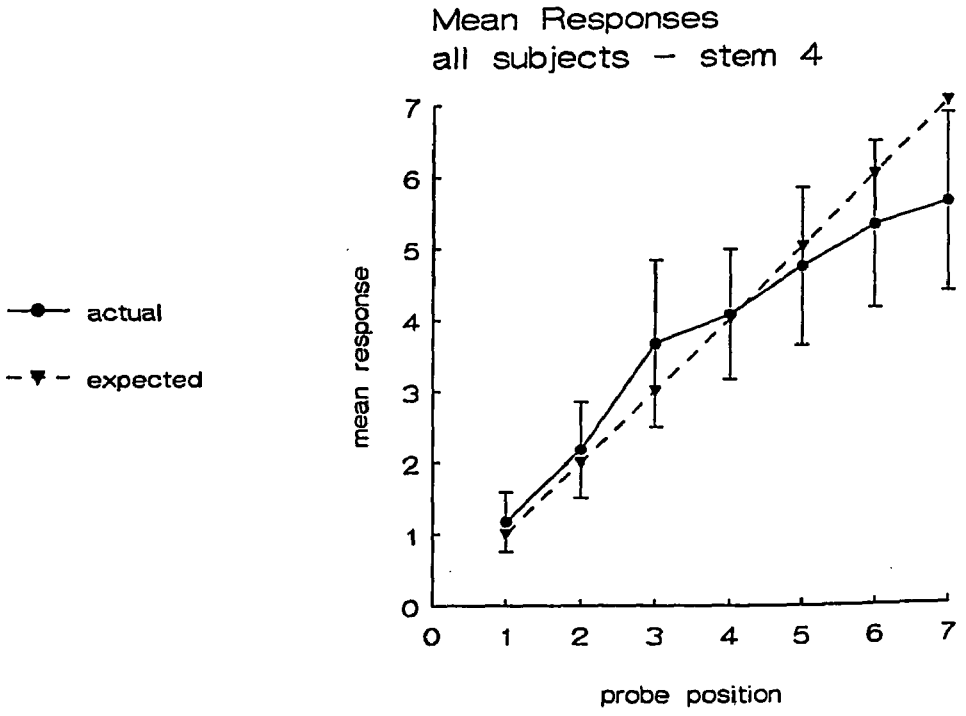


Fig. 2.3 Mean identification responses for all subjects. Error bars are standard deviations.

Figure 2.3 also shows that later positions are perceived as being earlier than they are; this replicates results from Sternberg and his colleagues (Sternberg, Knoll & Zukofsky, 1982) in a judgment task. Second, differences in proportions of correct responses from position to position are, on the whole, as predicted. Specifically, the positions with higher metric levels (1 or 2) tend to have higher proportions correct than their neighbors. An obvious exception is the high performance at position 1 (see below).

While these aspects of the data are consistent with BEATS' predictions, there are two arguments for a simpler response bias interpretation. First, the choices, in the pencil-and-paper task, corresponding to positions 2, 4 and 6 are ♪, ♩, and ♩., respectively. These note values occur more often in most music than do the values corresponding to positions 3, 5 and 7. Moreover, with both stems these values form sequences that are more likely to be familiar. Secondly, several subjects performed at or below chance (about 14% correct) at positions 3, 5 and 7. It seems quite plausible that subjects preferred responses corresponding to more familiar note values or rhythms. As shown in Table 2.4 and in Figures 2.4 and 2.5, the mean proportions of trials on which each response was used indicates that, with both stems, subjects did avoid

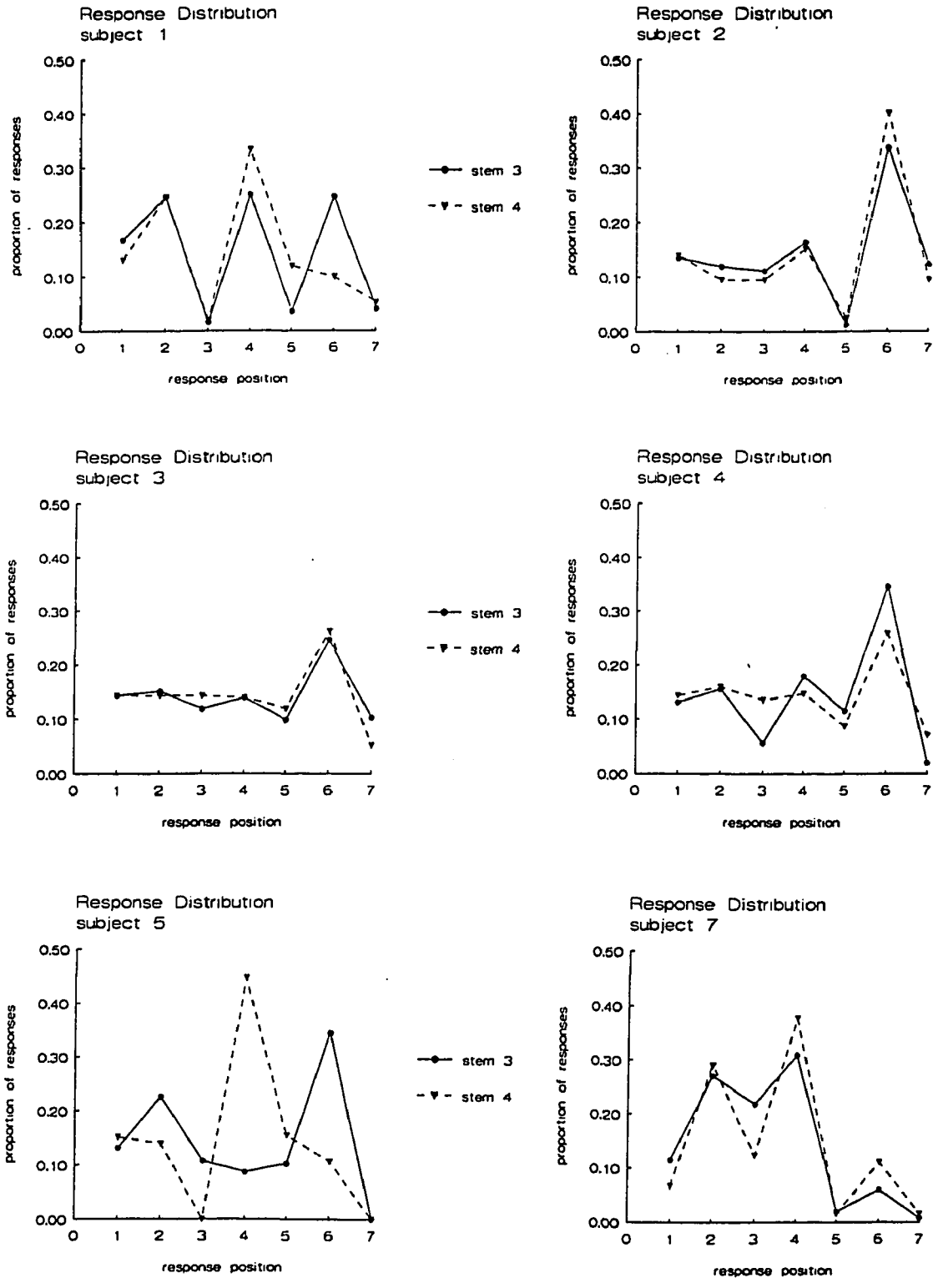


Fig. 2.4 Distributions of identification responses for each subject.

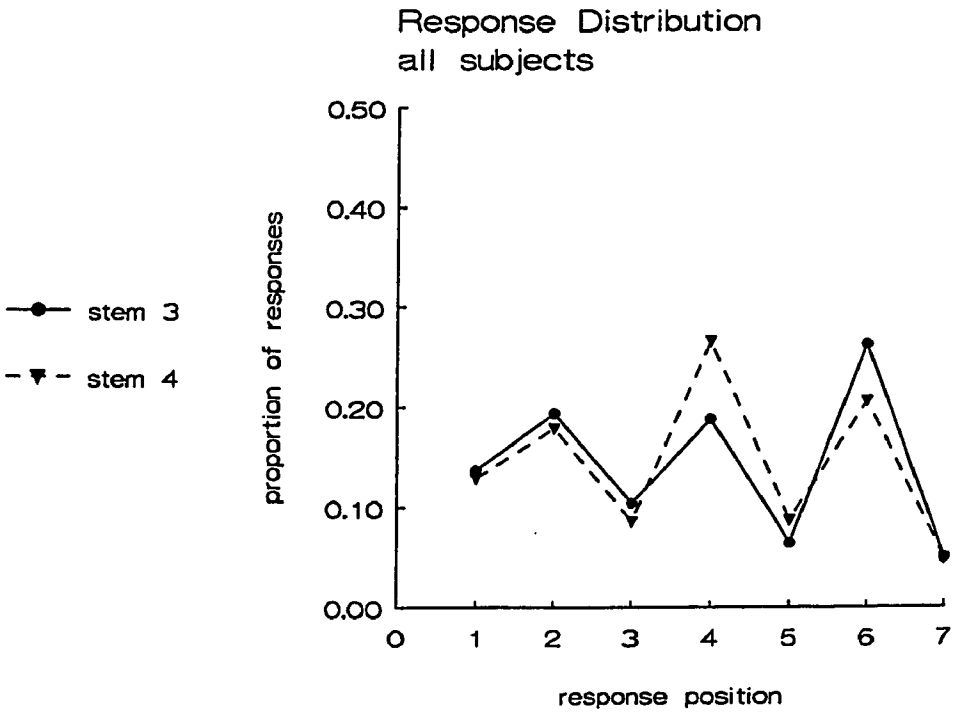


Fig. 2.5 Distribution of identification responses for all subjects.

responses 3, 5 and 7. Subject 5 avoided response 7 altogether.

	response						
	𝅘𝅥𝅯	𝅘𝅥𝅯	𝅘𝅥𝅯.	𝅘𝅥	𝅗𝅥𝅯𝅘𝅥𝅯	𝅗𝅥.	𝅗𝅥..
3-note stem	.14	.19	.10	.19	.06	.26	.05
4-note stem	.13	.19	.09	.27	.09	.21	.05

Table 2.4 Mean proportion of total responses for each stem.

The response bias interpretation is borne out by fitting a Thurstonian model (McNicol, 1972) to the identification data, pooled across subjects. Such a model represents the subject's task in signal-detection terms, and accounts for the data by establishing the distance between means (d') of the sensory distributions of each of the stimuli (in this case the seven probe positions) and by identifying the locations of the decision criteria.

	probe position						
	1	2	3	4	5	6	7
both stems							
d'		1.88	0.44	0.13	0.26	0.15	0.22
s.d.	1.0	0.40	0.46	0.44	0.55	0.54	0.63
3-note stem							
d'		1.91	0.27	0.22	0.25	0.08	0.29
s.d	1.0	0.38	0.46	0.42	0.57	0.50	0.68
4-note stem							
d'		1.88	0.69	0.10	0.28	0.30	0.30
s.d.	1.0	0.47	0.55	0.50	0.60	0.70	0.86

Table 2.5 d' values and standard deviations from the Thurstonian model of the identification data.

The d's are shown in Table 2.5, along with the standard

deviations of the sensory distributions. The standard deviations are expressed in terms of the standard deviation of the first distribution, which is arbitrarily defined as 1.0. The increasing dispersion of the sensory distributions of later probe positions is predicted by the Weber's law model proposed earlier. The d' values are represented graphically in Figures 2.6 and 2.7, which show the relative locations of the sensory distributions as well as the locations of the response criteria. The increasing variance and decreasing spacing of the sensory distributions account for the lower proportions of correct responses for the later positions (see Figure 2.2), while the locations of the criteria show strong response biases that account for the higher proportions of responses 2, 4 and 6 (see Figure 2.5). Table 2.5 and Figure 2.7 show that the sensory distributions associated with the four-note stem are somewhat more separated from one another. This may reflect more precise expectations regarding the probe note (Schulze, 1989) after hearing four notes as opposed to three. This greater separation accounts for the slightly higher performance with the longer stem, despite slightly higher sensory variance.

The Thurstonian model also explains the unpredicted high performance in position 1. The large dispersion of the sensory distribution for position 1 is offset by the

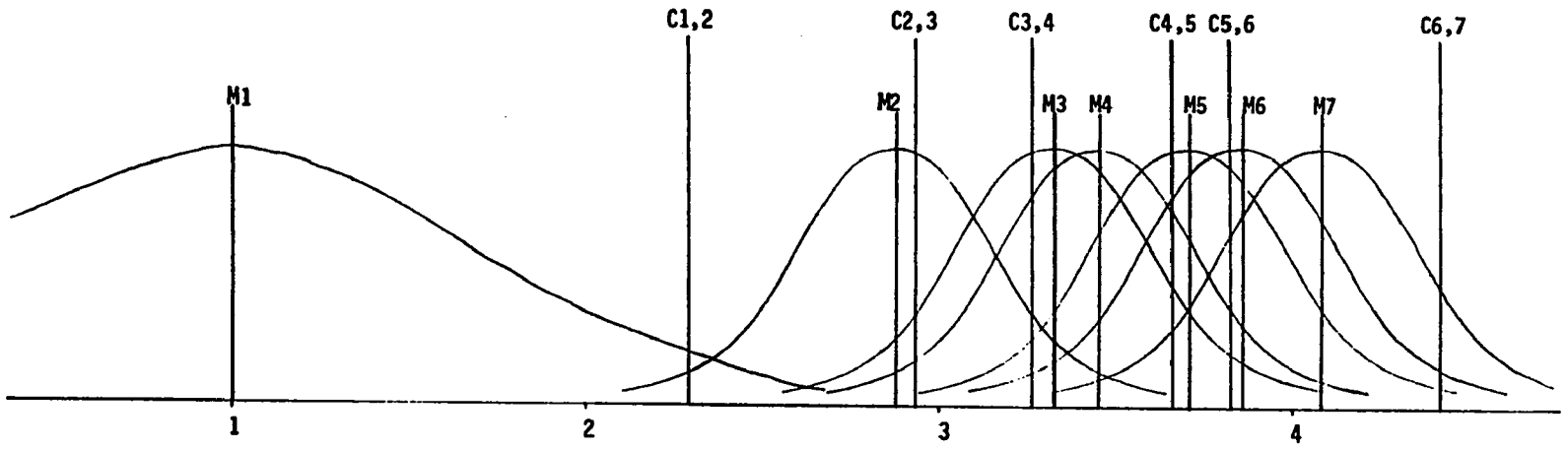


Fig. 2.6 A Thurstonian model of the identification data, showing the sensory distributions of the seven probe positions and the decision criteria, scaled in units of the first distribution.

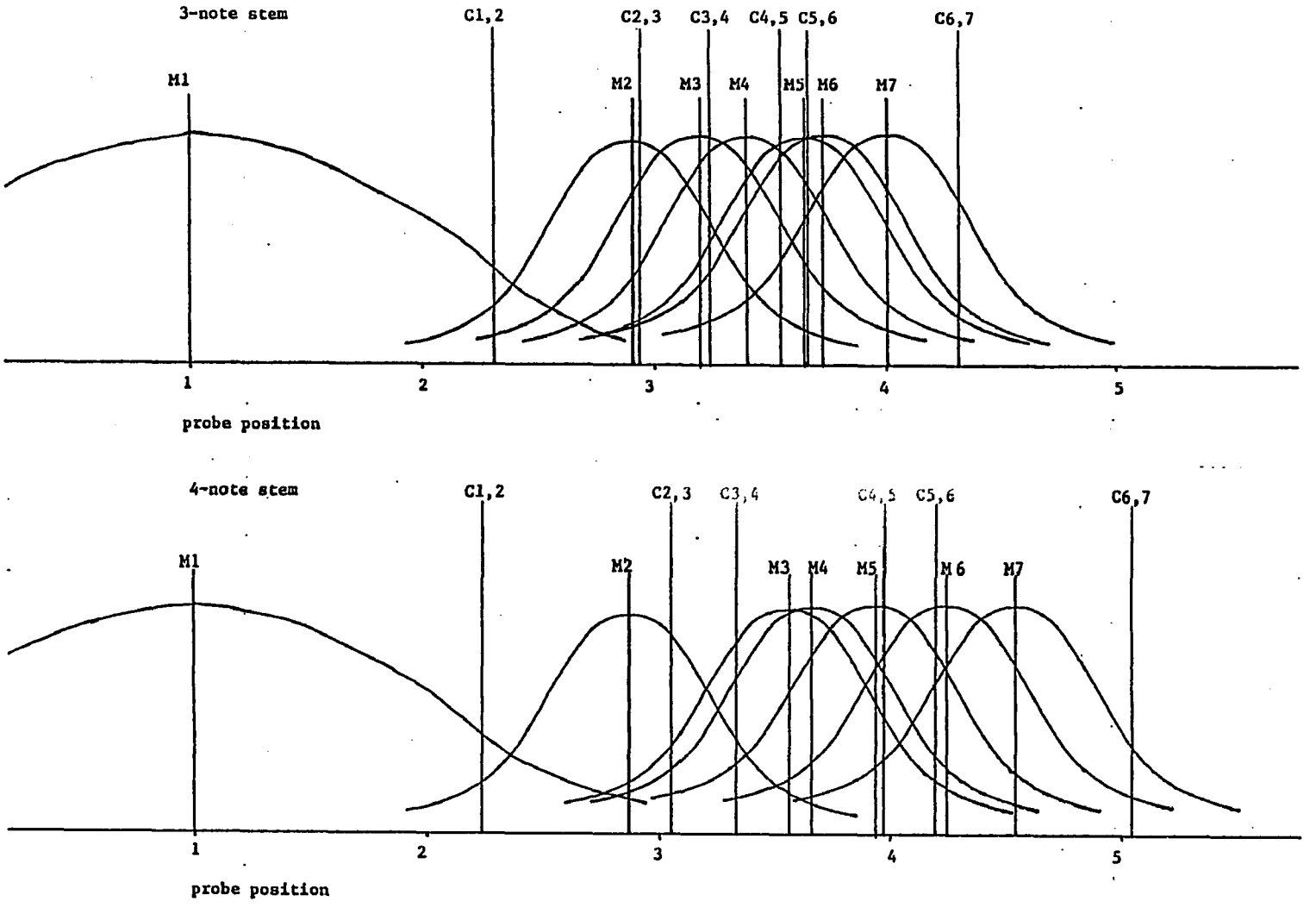


Fig. 2.7 Thurstonian models of the identification data, showing the sensory distributions of the seven probe positions and the decision criteria, scaled in units of the first distribution.

large d' for positions 1 and 2, which minimizes overlap of the two distributions. In addition, the criterion placement minimizes responses of "2" to stimuli in the first position. Finally, in those cases in which the probe is perceived to occur at an earlier position (i.e. in the left tail of the position 1 distribution), response 1 remains the earliest response the subject can make. Assuming subjects make the response that most closely approximates what they think they have heard, proportion correct will benefit from the fact that position 1 is the earliest position.

While response bias clearly had a large role in determining subjects' responses, there are aspects of the data that suggest that response bias is not the entire explanation. First, a response bias hypothesis predicts that the frequency distributions of responses will be the same for both stem lengths. However, the distributions differ from one another at response 4 (chi square = 26.2, $p < .01$) and at response 6 (chi square = 13.6, $p < .01$). The direction of the differences between distributions for these responses is consistent with the BEATS hypothesis. Second, the BEATS hypothesis predicts higher performance at higher metric levels. Thus for a given probe position performance should differ from one stem to the other if that position has a different metric level in each stem.

Response bias, on the other hand, assumes preference for responses on the basis of position rather than metric level, and therefore predicts no such difference in performance. At position 4, the three-note stem's probe is at metric level 1 while the four-note stem's probe is at level 2. Conversely, at position 6 the three-note stem's probe is at level 2 while the four-note stem's probe is at level 1. The resulting BEATS prediction is depicted below.

performance		pos 4	pos 6
stem	3-note	low	high
	4-note	high	low

Over the six subjects these four comparisons (each row and column above) yield differences in the predicted direction in 18 of 24 (6 subjects x 4 comparisons) cases; of these, six differences yielded significant correlated t values. While the overall decline in performance with later positions may account for the difference between positions 4 and 6 in the four-note stem, it cannot explain the same difference in the three-note stem. Likewise, neither overall decline nor response bias account for differences between stems at position 4 and at position 6.

These differences suggest that in addition to a pervasive pattern of response bias, the data reflect, to varying degrees, the operation of metric hierarchies as predicted by BEATS. A design that avoids confounding

BEATS' predictions with familiar note values and rhythmic figures should demonstrate more clearly the influence of metric hierarchies in identification tasks.

4. Experiment 3: Reproducing the Metric Probes

Subjects

Six musically trained subjects participated in six sessions apiece. Each session lasted approximately 30 minutes. None of these subjects had participated in either Experiment 1 or Experiment 2. These subjects ranged in age from late teens to early thirties; they included two pianists, a violinist, a guitarist, a flutist and a trumpeter. Five subjects had professional experience, and five had had training in rhythmic dictation.

Apparatus

Stimuli were presented, and responses collected, by a Turbo Pascal program, TPROBE, running on an AT-type computer. For stimulus presentation, TPROBE controlled the same apparatus used in the rating experiment. Subjects made reproduction responses by tapping on a MIDI drum pad (Roland MPD-4) connected via a MIDI interface (a common data protocol for linking electronic instruments and computers (Loy, 1985; MIDI Manufacturers Association, 1985)) to the computer. The drum pad was mounted inside a small covered wooden box. A wooden knob, connected by a

hinged lever to the box, protruded through a hole in the cover. A small steel ball was attached to the end of the lever directly under the knob, in such a way that pressure on the knob brought the ball into contact with the drum pad inside the box. When the knob was released a strip of elastic bearing on the lever returned the knob to its original position. The distance traveled between the knob's resting position and the surface of the drum pad was approximately two millimeters.

Procedure

Subjects were run in a sound-attenuating booth inside a small, closed room. Communication between the subject and the computer was through the small terminal in the subject's booth. The terminal's cursor was turned off to prevent its blinking from interfering with stimuli or responses.

Each session began with written instructions, followed by a series of practice trials. Subjects were told in advance that none of the practice stimuli would appear in the experiment. In the practice trials the subject was asked to reproduce rhythms by tapping on the drum pad. Subjects were not permitted to repeat an individual practice trial, but after the last practice trial they were given the option of repeating the practice session. The practice session could be repeated as often as the subject

liked. Subjects were encouraged to repeat the practice session until they were comfortable with the apparatus and the procedure.

After the practice trials subjects were asked if they had any questions, and then began the experimental trials. The trial cycle was designed to reduce subject fatigue by minimizing keystrokes and by pacing the experiment while allowing the subject to rest when necessary. The trial cycle was as follows:

1. Inter-stimulus interval. The computer paused. During this interval the subject could temporarily halt the program (to rest briefly, reposition terminal, chair, drum pad, etc.) by pressing the terminal's <ESC> key. This resulted in a message on the terminal instructing the subject to hit any key to resume. Alternatively, the subject could end the inter-stimulus interval immediately by pressing the space bar. Finally, the duration of the inter-stimulus interval could be toggled between long (two and a half seconds) and short (one second) by pressing the 'S' (slow) and 'F' (fast) keys during this interval.
2. Warning tone. The terminal beeped. The computer did not respond to any of the keys mentioned above after this.
3. Attention interval. Following the warning tone was a pause for a random period in the range 1000 - 1400 milliseconds. Any duration in this range is longer than

the first interval of any of the stimuli. This, in conjunction with the fact that the warning tone is considerably higher pitched than the stimulus tone, prevented the warning tone from becoming part of the stimulus.

4. Stimulus presentation.

5. Response. The computer waited for the subject to tap.

6. End of response. A tapping response was considered finished when, after the subject had begun tapping, two seconds passed without a tap. This was done to allow the subject to keep his/her hand on the drum pad between trials.

Results and Discussion

All dependent variables used in these analyses are based on the tapping error, i.e. the difference between the last interval of the stimulus and the last interval tapped by the subject. Proportional errors (i.e. the tapping error expressed as an unsigned proportion of the final stimulus interval) for each subject are shown in Figure 3.1 (three-note stem) and Figure 3.2 (four-note stem), and summarized in Figure 3.3. Mean proportional errors for each stem x level combination, pooled across positions, subjects and tempos, are shown in Table 3.1.

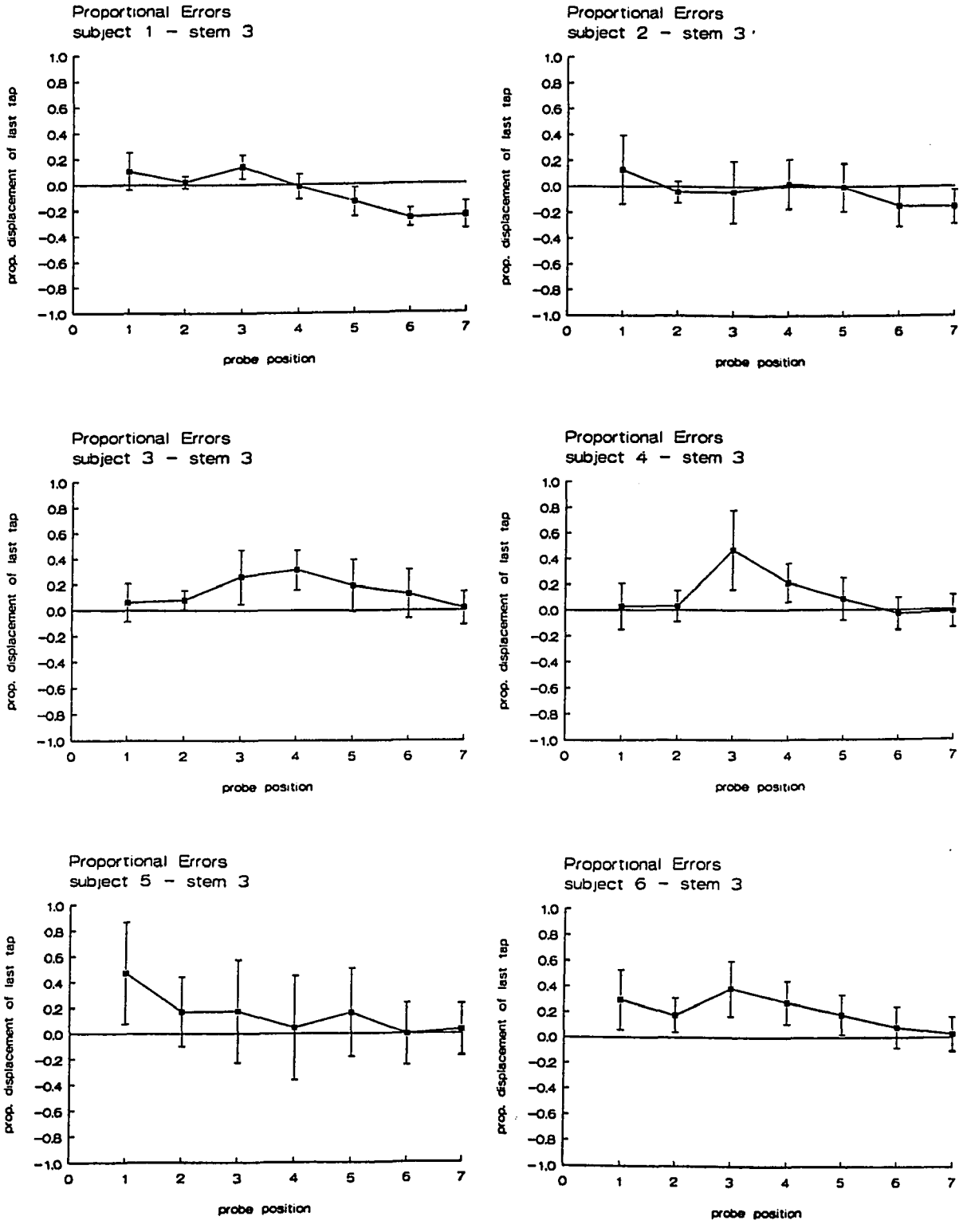


Fig. 3.1 Proportional errors for each subject for three-note stem in the reproduction task. Error bars are standard deviations.

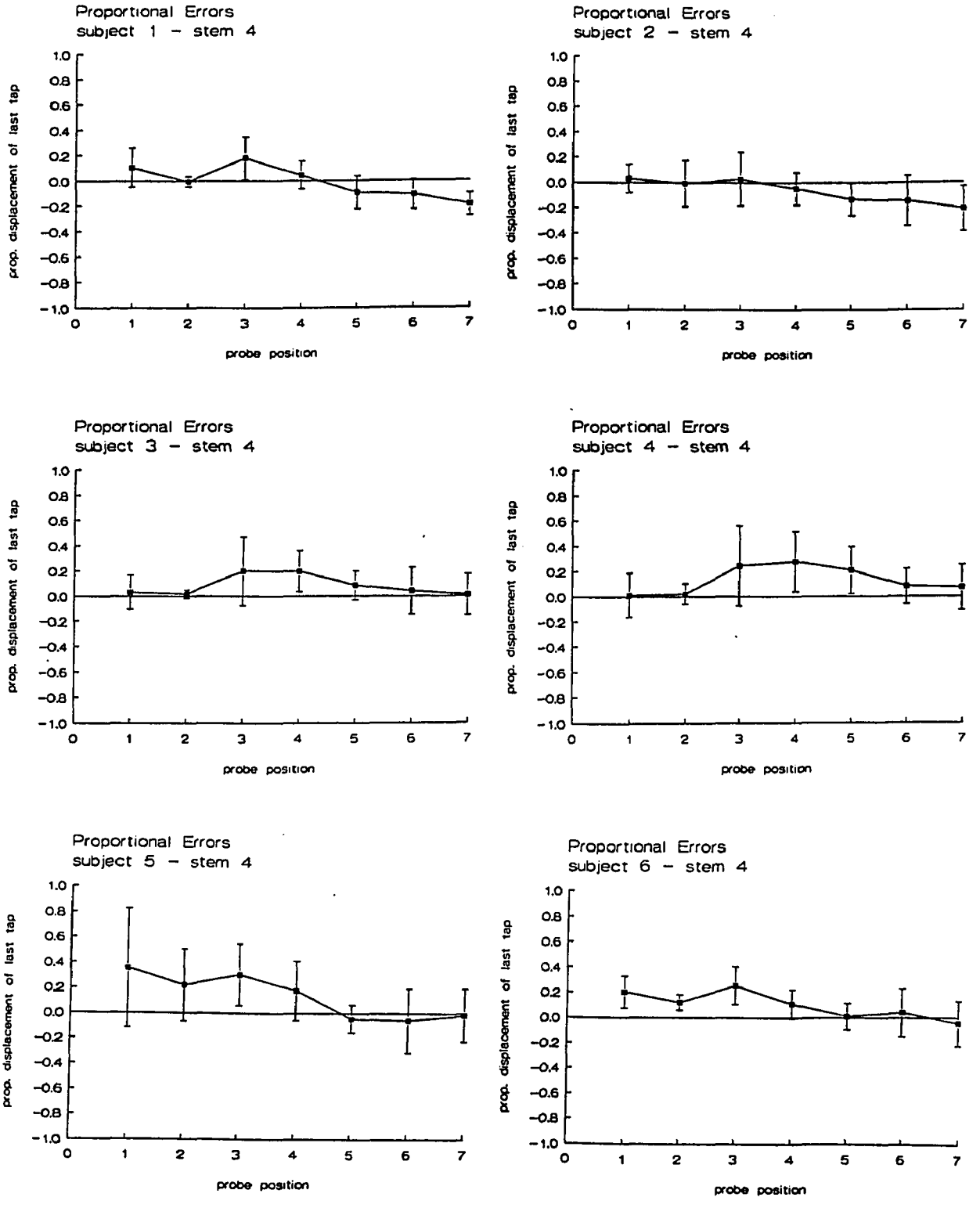
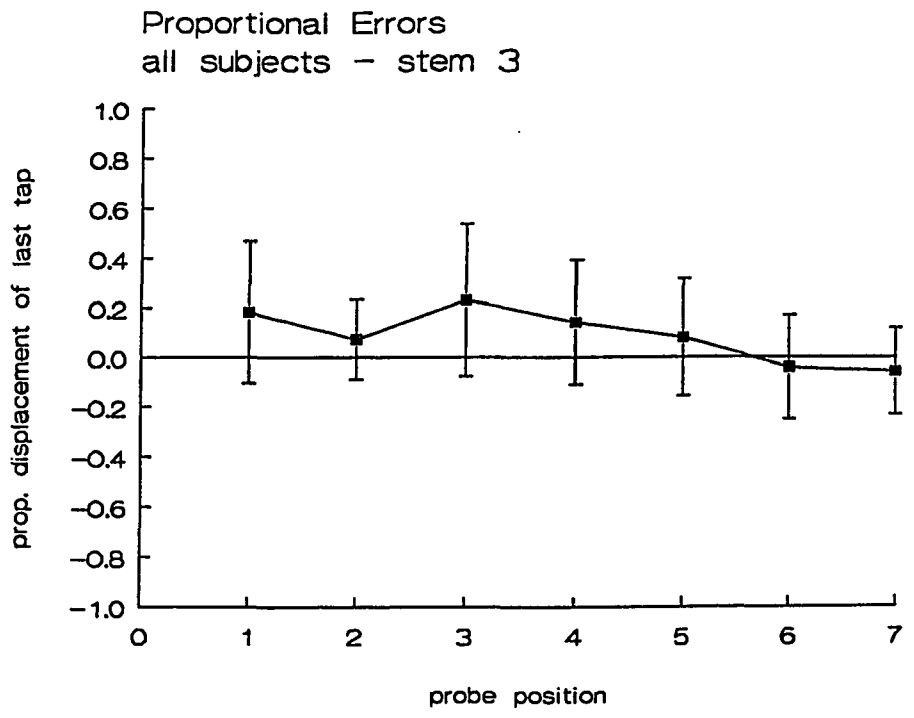


Fig. 3.2 Proportional errors for each subject for four-note stem in the reproduction task. Error bars are standard deviations.



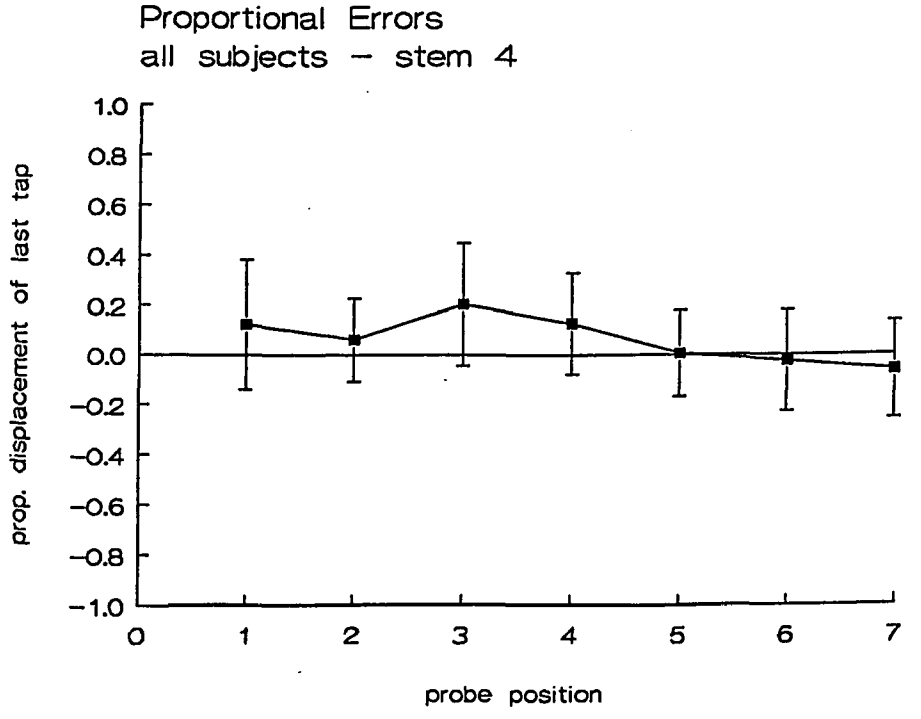


Fig. 3.3 Mean proportional errors in the reproduction task. Error bars are standard deviations.

	metric level 0	1	2	mean
stem length: 3	0.25	0.25	0.16	0.22
4	0.20	0.15	0.18	0.18
mean	0.22	0.18	0.17	0.20

Table 3.1 Mean proportional reproduction errors by metric level.

Preliminary analysis showed no effects of session or stimulus order, and these factors were eliminated from further analysis. Mean proportional tapping errors for each subject were analyzed in a 2(stem) x 7(position) x 3(tempo) analysis of variance design with repeated measures on all factors; a second analysis replaced the position factor with the level factor. There is a significant effect of stem length ($F_{1,8} = 8.0$, $p < .05$). As shown above, the direction of the difference (smaller errors for the longer stem) is consistent with the higher ratings and the higher proportion of correct responses for the longer stem found in the first two experiments. The effect of metric level is also significant ($F_{2,10} = 6.0$, $p < .05$); as shown above, errors are smaller for higher levels. There are also significant interactions of position x tempo ($F_{12,60} = 3.14$, $p < .01$) and of stem length x position x tempo ($F_{12,60} = 3.0$, $p < .01$).

There is a significant effect of probe position ($F_{6,30} = 5.58$, $p < .01$). Mean proportional errors for each stem x position combination, pooled across subjects and tempos, are shown in Table 3.2.

mean prop. error stem length	probe position 1	2	3	4	5	6	7
3	.24	.12	.34	.25	.22	.20	.18
	(0)	(2)	(0)	(1)	(0)	(2)	(0)
4	.19	.11	.27	.18	.16	.19	.19
	(0)	(1)	(0)	(2)	(0)	(1)	(0)

(metric level in ( ))

Table 3.2 Mean proportional reproduction errors by probe position.

The probe interval can be expressed as a time interval (which depends on tempo), as a note value (e.g. ♪, ♩, 𝅗𝅥, which is independent of tempo), or in terms of the metric level of a note value (which is also independent of tempo). The BEATS hypothesis predicts that metric level will give the simplest account of the data. By contrast, a simple time perception/production hypothesis predicts that time intervals will do the best job. A third hypothesis might take an intermediate approach, predicting that reproduction accuracy will be a function of the probe interval's note value regardless of the time interval. These hypotheses can be distinguished by their predictions regarding 12 pairs of stimuli (six for each stem) in which the two probe intervals represent the same time interval but different note values and in some cases different metric levels. These pairs are listed in Table 3.3, together with mean absolute errors. T-tests ($\alpha = .05$) were used to compare these means for each stimulus pair; significant differences

are indicated by < and >, non-significant differences by ≤ and ≥.

stem	-- stimulus 1/2 -- note	level	time	mean absolute error (msec.) stim 1		stim 2	-- hypotheses -- time	note	metric
3	♪/♪	0/2	200/200	47.1	>	18.1	=	<	>
4	♪/♪	0/1	200/200	33.6	≥	15.5	=	<	>
3	♪/♫	2/0	300/300	40.2	<	109.7	=	<	<
4	♪/♫	1/0	300/300	29.0	<	80.1	=	<	<
3	♪/♩	2/1	400/400	51.8	<	95.1	=	<	<
4	♪/♩	1/2	400/400	63.6	≤	94.4	=	<	>
3	♫/♩.	0/2	600/600	195.5	≥	126.6	=	<	>
4	♫/♩.	0/1	600/600	131.3	≥	128.3	=	<	>
3	♩♪/♩.	0/0	700/700	174.4	≥	147.9	=	<	=
4	♩♪/♩.	0/0	700/700	93.0	<	143.5	=	<	=
3	♩♪/♩.	0/0	1000/1000	196.7	≥	182.2	=	<	=
4	♩♪/♩.	0/0	1000/1000	147.1	≤	156.0	=	<	=

Table 3.3 Contrasts of stimulus pairs with the same probe interval.

Table 3.3 also shows the direction of the difference between means predicted by each hypothesis. For example, in the first pair the mean error for the first stimulus is significantly greater than the mean error for the second stimulus, as predicted by the metric level hypothesis. In the second pair, on the other hand, there is no significant difference, as predicted by the time interval hypothesis, but the direction of the difference is that predicted by the metric level hypothesis. The time interval hypothesis makes the same prediction in all cases, namely that the errors should not differ significantly. Likewise the note

value hypothesis makes the same prediction in all cases, namely that errors should be smaller for smaller note values. The metric level hypothesis predicts that the probe with the higher metric level should have the smaller error, and that when both probes have the same metric level their errors should not differ significantly. The above table does not unambiguously rule out any of these hypotheses. There are seven cases in which the time interval hypothesis cannot be rejected, but five in which it can. While the other hypotheses cannot for the most part be framed as null hypotheses, their predictions can be compared to the direction of the difference between means. The note value hypothesis predicts four of the five significant differences, and over all agrees with five of the mean differences. The metric level hypothesis predicts four of the significant differences, and over all agrees with ten of the mean differences. The note value and metric level hypotheses can be cast as null hypotheses by contrasting mean errors for stimulus pairs having the same probe position (and hence the same note value and metric level) but differing in tempo. Comparing means across each of the three pairs of tempos at each of the seven probe positions yields 21 comparisons for each stem. T-tests (α = .05) were used to compare these means for each stimulus pair and are summarized in Table 3.4; significant

differences are indicated by < and >, non-significant differences by ≤ and ≥.

relation of long to short	frequency	predicted by
<	1	---
≤	5	note, metric
≥	22	note, metric
>	14	time

Table 3.4 Summary of contrasts of stimulus pairs with the same probe position.

Three of the significant differences in Table 3.3 are predicted by both the note value and metric level hypotheses. Inasmuch as note value determines metric level, and metric level does not determine note value, the most economical interpretation of these three cases is to suppose that they are predicted by the note value hypothesis only in the sense that the note value hypothesis makes the same predictions as the metric level hypothesis. While the note value hypothesis perhaps cannot be ruled out entirely, a coherent general account of the data can be made in terms of time intervals and metric levels. Other things being equal (as in Table 3.4), the larger the time interval, the larger the error. At the same time, the higher the metric level, the smaller the error. Both predictions are consistent with a Weber's law model like that of the rating task. By contrast, note value predicts errors only to the extent that it is redundant with metric level.

It is interesting to compare this analysis of the reproduction data with an experiment by Sternberg and his colleagues (Sternberg et al., 1982, experiment 9). In the present terminology, their experiment was an attempt to determine whether a time interval hypothesis or a note value hypothesis gave a better account of subjects' performance in a judgment task. Subjects heard a pair of clicks, defining an interval, followed by a third click separated from the second click by a fraction of the initial interval. The fraction was varied from trial to trial, using an adaptive procedure, around four target fractions (1/8, 1/6, 1/4 or 1/2). Subjects' task was to judge whether the presented fraction was larger or smaller than the target fraction. The resulting psychometric functions defined mean fractions (note values) and time intervals associated with each target. Comparison of the same fractions across initial interval durations (tempos) allowed a test of the note value hypothesis, while comparison of the same durations across fractions allowed a test of the time interval hypothesis. Both hypotheses were rejected, though the results were slightly more consistent with the time interval hypothesis than with the note value hypothesis.

There was no metric level hypothesis in the Sternberg et al. experiment because the single interval that preceded

the interval to be judged could not be expected to induce a metric hierarchy in the listener's representation of the stimulus. This difference notwithstanding, their results are in agreement with the present data in the sense that in neither case is there support for a time interval hypothesis alone or for a note value (or metric level) hypothesis alone. Sternberg et al. do not suggest what kind of hypothesis might give a better account of their data, but it is clear that they do not consider either time interval or note duration to be irrelevant simply because each considered in isolation is inadequate. It is possible that, as suggested above in the context of the present experiment, a model combining time interval and note duration would provide a better explanation of Sternberg et al.'s data.

It is clear that these data do not unambiguously support the BEATS hypothesis. However, one general finding appears quite robust, which is that the relation between probe position and tapping error, however measured, is nonmonotonic. This is not to say that time intervals and note values have no predictive value. Rather, it suggests that a successful model will be one like that suggested in the context of Experiment 1, in which monotonic and nonmonotonic structures interact.

5. Summary of the Probe Experiments

The three probe experiments agree in their support of the hypothesis that meter is represented hierarchically. While the support drawn from any one experiment is diminished variously by high variability, low sensitivity, response bias, and perhaps other factors, the fact that all three point, however tentatively, in the same direction can be construed as additional support for the hypothesis.

There are two general implications of these experiments that are worth emphasizing. First, meter can be induced by rhythms composed of note durations only. This does not mean that duration is the sole cue to meter, but it does mean that listeners have perceptual strategies for extracting meter from duration, which they use, at the very least, when there are no other cues. The conditions under which such strategies do and do not come into play cannot be extrapolated from the present experiments. However, subjects apparently extracted meters without being explicitly instructed to do so, in spite of the fact that extracting meter from short monotone duration sequences is not a common musical task. This suggests that the meter extraction strategies exhibited in these experiments are not reserved for duration-only contexts, but rather are part of a larger repertory of perceptual skills relevant to music.

Second, meter can be induced quickly. The stimuli in these experiments could not have been much shorter, yet by the fourth or fifth note subjects were able to distinguish two and perhaps three distinct metric levels. This is not to say that the process ends here, but it suggests that, like BEATS, listeners have some efficient strategies that are able to make use of small amounts of evidence.

An equally important conclusion to be drawn, particularly from the rating and reproduction experiments, concerns the interaction of time and metric structure. Models of music perception cannot treat meter as an ideal structure, independent of the sensory processes that subserve it. Meter should not be thought of as an unflagging metronome (or set thereof), but rather as a set of predictions about the next few events. These experiments demonstrate that a metric structure degrades rapidly, even over a short span of time, when there are no notes. This may help to explain why, in the overwhelming majority of music, there are plenty of notes and silences are short and few. By extension, sparse music with many long silences ought to be difficult to comprehend, and by most accounts it is. Learning to listen to such music may be, in part, a matter of learning to do without meter.

An important qualification of the above remarks is that it may well be that a metric structure that has carried its

listener halfway through a long piece may be more resistant to decay than one that is only four or five notes old. Schulze's (1978) results from isochronous sequences, as well as the slightly higher ratings for the longer stem in Experiment 1, suggest that there is at least a small effect in this direction, but studies with longer musical stimuli are needed to determine the nature of this relation.

6. Experiment 4: Error Matching

While BEATS would be untenable without evidence that listeners do in fact extract hierarchical representations of meter, such evidence does not by itself support BEATS as a model of the process by which listeners perform such extraction. Experiment 4 addressed this problem.

Rationale

A testable prediction comes from the fact that BEATS often fails where people succeed. For example, in a random selection of 248 folk songs (Lomax, 1960), BEATS was completely correct in only 95 cases (38%). That BEATS, using timing information only, should often fail where human listeners succeed is not surprising, given that the latter can take advantage of other cues, such as melody, accents, and harmony. But do humans do as well when their judgments are based, as BEATS' are, on timing information alone? If BEATS is a plausible model of time-based meter

perception, human listeners should succeed where BEATS does and make the same errors where BEATS fails. To test this prediction, BEATS' performance was compared with that of musically skilled listeners over a range of music samples. These samples included a variety of scores that BEATS analyzed successfully as well as scores that led to each of BEATS' characteristic ways of failing (to be described).

In order to compare BEATS' performance to that of subjects, it is necessary to elicit a response that provides enough detail about the subject's metric interpretation of a given piece to make a meaningful comparison with BEATS' analysis. Asking subjects to determine the time signature provides some indication of the relation between levels in the metric hierarchy, but it does not indicate the perceived phase relation between the metric hierarchy and the rhythm. The subjects' task in this experiment was two-handed tapping, with the two hands tapping at two different rates in a ratio of 2:1 or 3:1. One-handed tapping reveals a single level of the listener's metric hierarchy, but allows no inferences about the hierarchy itself. By adding a second hand, the tapping listener reveals a second level, which indicates the phase of the metric hierarchy as well as its organization (duple or triple). While two levels may not be the whole hierarchy, two levels and their relations to the score

sufficiently characterize a hierarchy that useful comparisons can usually be made to the hierarchy produced by BEATS. Subjects tapped to a given score at each of several tempos. In general, subjects should tap higher metric levels (i.e. larger note values) at faster tempos and lower metric levels at slower tempos (Fraisse, 1982; Handel, 1984), thus providing a more complete picture of the metric hierarchy.

There is experimental evidence suggesting that, in addition to being useful in this experiment, two-handed tapping makes sense to subjects. For example, Deutsch (1983) asked subjects to divide an isochronous pulse train into 1, 2, 3, 4 or 5 intervals with each hand. The five possibilities for each hand were combined factorially so that all combinations were recorded. Subjects performed relatively poorly when asked to tap non-integral combinations (e.g. 3 against 2). Deutsch takes this to be evidence of an integrated, hierarchical representation of the pulse trains tapped by the two hands. Integral combinations can be represented by simple hierarchies, whereas non-integral combinations are represented by more complex hierarchies (or perhaps by no hierarchy at all). Ease of hierarchical representation results in greater timing accuracy. This is in keeping with an earlier analysis of timing accuracy in tapping by Vorberg & Hambuch

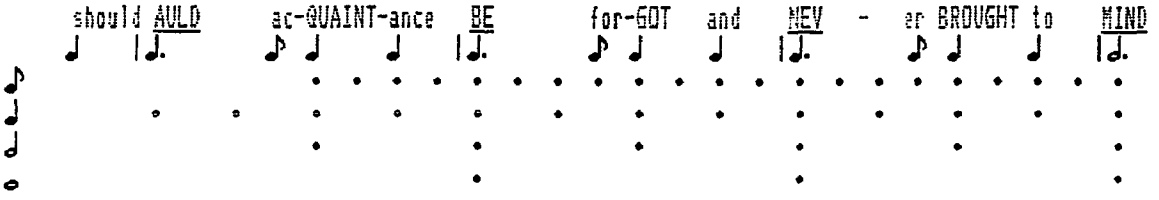
(1978), who found support for a model in which timing is carried out by simple chained timekeepers, but under the control of a hierarchical motor program. The fact that two-handed tapping itself seems to be represented hierarchically makes it a natural choice for this kind of experiment.

Classifying BEATS' Errors

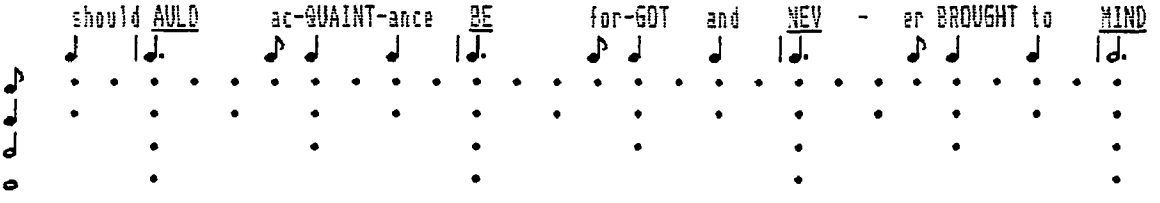
When BEATS' analysis of a given score is compared with a pencil-and-paper analysis (henceforth standard) of the full score (i.e. including melody, lyrics (if any), and other markings), the correspondence falls into one of several categories.

1) Agreement between BEATS and standard, as in "Auld Lang Syne:"

BEATS:



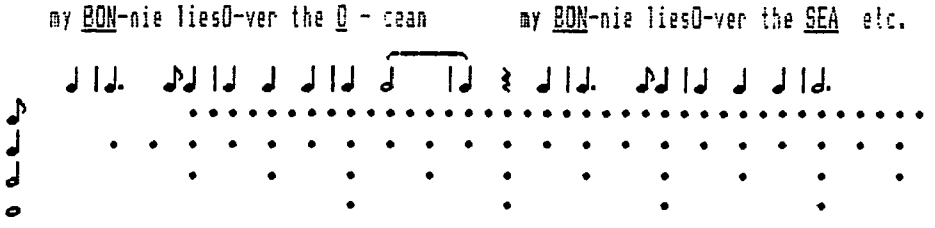
Standard:



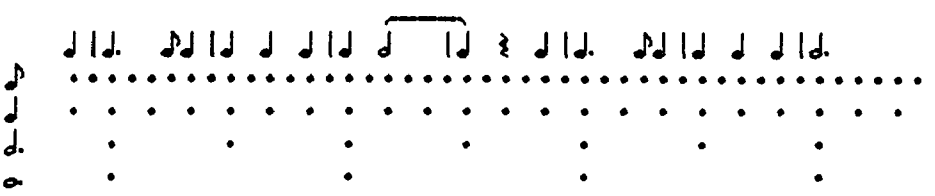
In this example BEATS generates the same levels, in the same phase, as are found in the standard.

2) Disagreement regarding meter, as in "My Bonnie:"

BEATS:



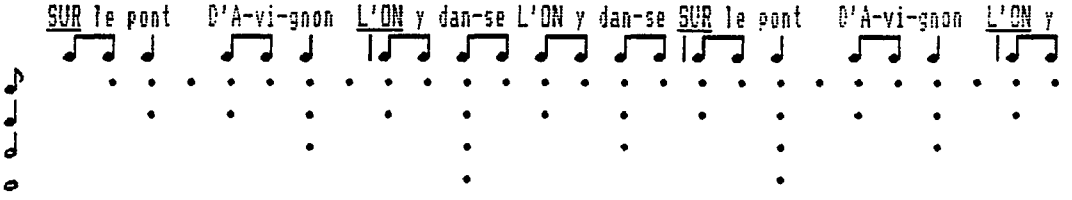
Standard:



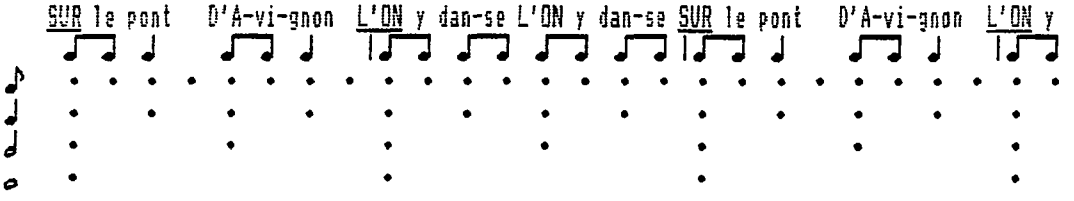
Here BEATS has found a duple meter (4/4), whereas the standard shows a triple meter (3/4). The critical level is the ♩ level, which groups ♩s by twos rather than threes.

3) Disagreement regarding phase, as in "Sur le Pont d'Avignon":

BEATS:

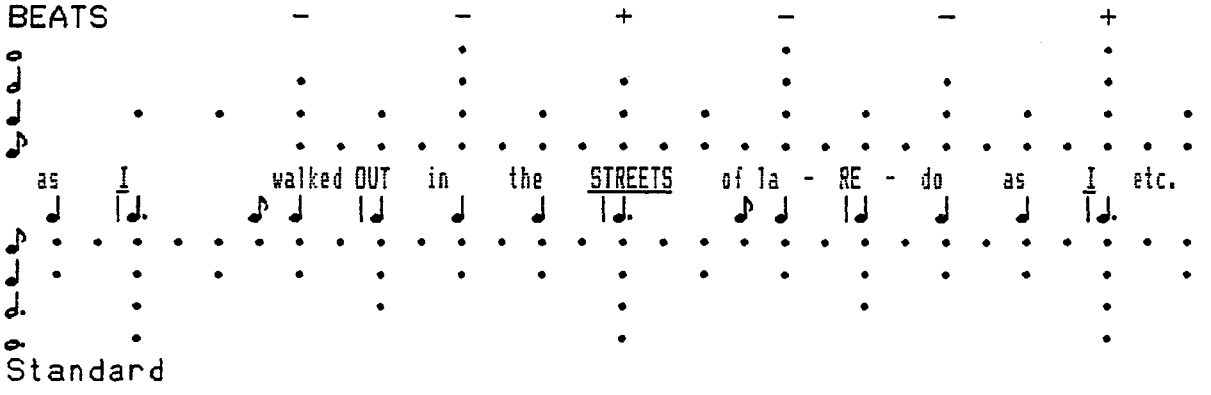


Standard:



Here BEATS has generated the same levels that are in the standard, but their phase is off: BEATS would sing "sur le PONT d'aviGNON," etc.

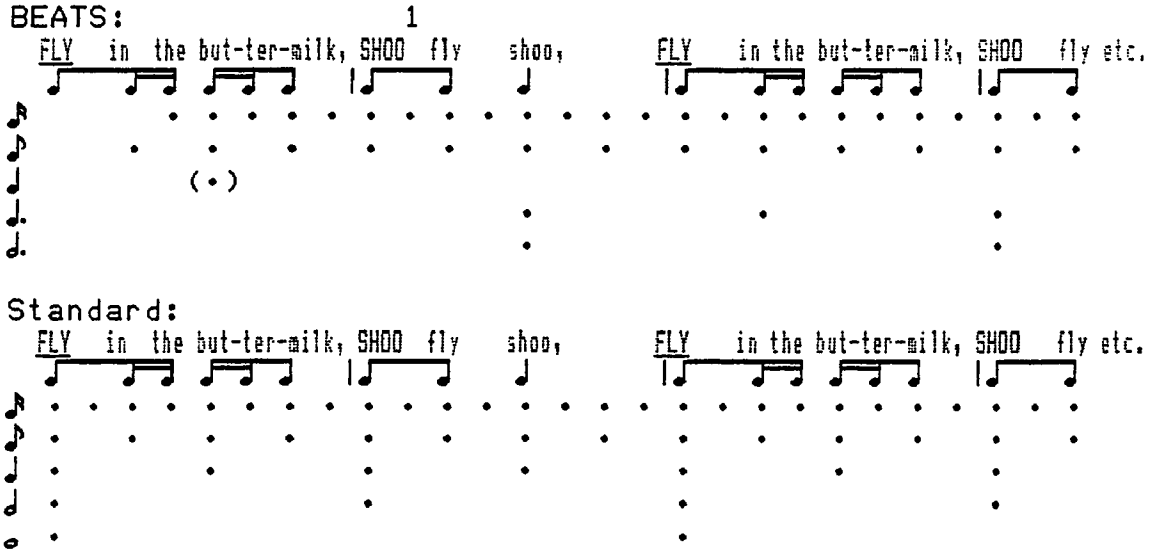
Note that meter and phase errors usually cannot be combined in any meaningful way. If the meter is incorrect, it makes no sense to ask whether the metric hierarchy is in phase. Any incorrect metric level will periodically generate a strong beat at the right place in the music. For example, in BEATS' analysis of "Streets of Laredo" (below), the ♪ and ♩ levels match the standard, but the ♩ and 𝅗𝅥 levels create strong beats that are alternately in (+) and out (−) of phase with respect to the corresponding ♩. and 𝅗𝅥. levels of the standard:



The only meaningful combination of meter and phase errors is a hierarchy in which a lower metric level is metrically incorrect (i.e. does not belong to the standard hierarchy) and a higher metric level is metrically correct but out of phase. This was true of only one stimulus and did not warrant a separate category. The combination does, however, occur in subjects' tapping responses (see below).

Meter errors fall into two distinct categories. The first, exemplified by the analysis of "My Bonnie" (above),

might be called the default meter error: from the beginning BEATS generates a well-formed hierarchy that contains (an) incorrect level(s). The second type is the stretch error, in which a correct analysis is spoiled by a Stretch operation which occurs late enough that it is clear that the analysis was correct up to the application of Stretch. An example of this type of error is BEATS' analysis of "Skip to my Lou:"



BEATS' analysis is correct at the outset, with ♬, ♪ and ♩ levels. At the second "shoo" BEATS would Double the ♩ level to produce a 𝅗𝅥 level. At location 1, however, the conditions for the Stretch operation (see above) are met, and BEATS generates a 𝅗𝅥. level, which is subsequently Doubled to produce a 𝅗𝅥. level. Since the ♩ and 𝅗𝅥. levels conflict, one of them must be removed. In this case BEATS estimates the meter to be triple (see Remove rule in

Appendix A) and finds the ♩. level more appropriate than the ♩ level. BEATS therefore removes the ♩ level. The triple meter hierarchy that BEATS constructs is well-formed but incorrect; the Stretch operation has spoiled what would otherwise have been a correct analysis. Because the above scenario is quite distinct from the way a default error is made, it seems reasonable to distinguish between default errors and stretch errors in this experiment. In some cases BEATS is on its way to a default meter error when a Stretch produces the upheaval described above, resulting in a different but still incorrect analysis. Because it is not clear how such cases ought to be classified, they have been excluded from the experiment.

There is a roughly corresponding typology of phase errors. The first occurs when a score has an upbeat or upbeats but the conditions for the Upbeat or Slide operations are not met. BEATS proceeds to construct a metric hierarchy that treats the first note as a downbeat. This type of error is called a negative phase error, reflecting the fact that the error is due to BEATS failure to apply the appropriate rule. The second type of error, the positive phase error, is the mirror image of the negative phase error. Here a score that begins on a downbeat happens to meet the conditions for the Upbeat or Slide operation, with the result that BEATS treats the

first note as an upbeat.

The classification of errors presented here is based not only on results but on processes as well. Looking at a score and a metric structure that represents a meter error, there is no way to know whether the error is due to BEATS' default operation or to a misapplication of the Stretch operation. Likewise, negative and positive phase errors look the same. In both cases the distinction is based on a difference in the process(es) responsible for the error. These distinctions allow a more detailed evaluation of BEATS. For example, if subjects make meter errors when BEATS makes default errors but not when BEATS makes stretch errors, it might be inferred that the conditions for the Stretch operation are not restrictive enough. A corresponding asymmetry between negative and positive phase errors would permit similarly detailed inferences.

Subjects

Sixteen musically trained subjects participated in four one-hour sessions each. Some of these subjects had participated in one of the metric probe experiments described earlier. Subjects ranged in age from late teens to mid-forties, with most in their twenties. They included four guitarists, three singers, two pianists, two string players, two clarinettists, two brass players, a flutist and a percussionist. Subjects had from three to twenty

years of training, with a mean of 11.8 years. Subjects were paid $5 per session.

Stimuli

It was important that subjects not recognize the stimuli, as this would change the nature of the task from perception to performance. White (1960) found 33% recognition of duration-only stimuli using a set of very familiar songs. Accordingly, scores for this experiment were selected from a large body of Anglo/American folk songs (Sandburg, 1927; Lomax, 1960; Warner, 1984), most of which were judged likely to be unfamiliar to most subjects. There are several reasons for using folk songs rather than, say, Mozart symphonies or Bartok quartets. First, songs are naturally monophonic. A listener who extracts a meter from a song extracts it from a single voice, whereas in the case of ensemble music it would be difficult to identify the respective contributions of the several voices. To use ensemble music one would have to be able to demonstrate that the single voice presented yields an unambiguous meter. Second, folk songs are rarely metrically ambiguous; the extraction of meter from folk songs is generally successful and effortless. Third, folk songs rarely change tempo or meter, and in any case scores that do were not used in this experiment. Finally, folk songs are short; using the whole song is a more satisfactory arrangement

than having to decide how much of which voice from an ensemble piece to use.

Stimuli were selected by using BEATS to analyze a large number of folk songs. Each BEATS analysis was compared to the score and classified as described above. The resulting distribution, broken down by time signature[1], is shown in Table 4.1.

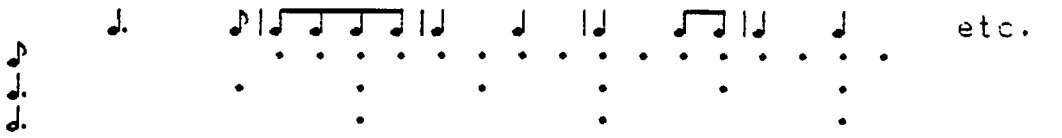
meter	BEATS error category correct	default	stretch	+ phase	- phase	
2/4	19	2	14	5	13	53
4/4	42	4	28	8	21	103
3/4	34	50	1	2	0	86
6/8	21	29	1	0	5	55
	116	85	44	15	39	297

Table 4.1 Distribution of BEATS' errors by time signature.

There are several important disparities in this distribution. First, there are relatively few default errors in duple-meter scores. In a default error, the meter error is not due to the (inappropriate) application of the Stretch operation but rather to the failure of the Stretch operation to apply. In triple-meter scores this is common: Stretch fails to apply and Double blithely builds a well-formed but incorrect duple metric hierarchy. An

1. In the sample, four time signatures (duple meters 2/4 and 4/4 and triple meters 3/4 and 6/8) accounted for 96% of the scores analyzed. The other time signatures (2/2 and 3/8) were not used in the experiment.

equivalent error can occur in a duple-meter score only when the first note is dotted



and there is no basis for Stretch to enlarge the metric unit to an appropriate size. Here, the initial ♩. establishes a metric unit of ♩.; for BEATS to construct a duple metric hierarchy, it would need to enlarge this metric unit to ♩, in effect combining ♩. and ♪ in a larger unit. Instead, Double applies, creating the 𝅗𝅥. metric level and leading to classification as a default error.

The second disparity is the mirror image of the first, i.e. the relative scarcity of stretch errors in triple-meter scores. In a stretch error, an application of the Stretch operation enlarges an appropriate metric unit to an inappropriate size. In duple-meter scores this is common and results in a triple metric hierarchy. In triple-meter scores, however, Stretch is usually required to discover the triple meter; when this happens the score is, other things being equal, correctly analyzed. For a stretch error to occur in a triple-meter score, the Stretch operation must either yield an incorrect metric unit (as when early note durations are distorted by rests or ties) or be unnecessary in the first place (as in scores that begin with dotted notes). For whatever reason, this almost

never happens.

The third disparity, somewhat less pronounced than the first two, is the relative infrequency of phase errors in triple-meter scores. Many scores that might end up as phase errors are classified as default meter errors. A meter error has priority over a phase error in the sense that the meter must be right before it makes sense to ask if the phase is right. The small number of triple meter phase errors is probably accounted for by the higher rate of meter errors among triple-meter scores (57% as opposed to 31% for duple-meter scores).

The design called for four scores to be selected from each of the cells in the above distribution. Because the folk-song rhythms did not fill all the cells, approximately 100 new rhythms were created in order to complete the design. Each of these rhythms was conceived as a simple eight-bar tonal melody in a given time signature. This was intended to insure that the resulting rhythms would make musical sense and that the standard analyses of these rhythms would not be arbitrary. The new rhythms were analyzed by BEATS, and those that fell into needed error categories were added to the pool of potential stimuli, providing enough of the desired types of errors to allow stimuli to be chosen from each error/meter category. When there were more than enough cases in each cell, scores were

selected randomly.

Stimuli replicated the conditions in which BEATS operates: all notes had the same pitch, intensity, articulation and timbre, with note duration the only cue to meter. The tone sounded for the initial three-fourths of the note duration. All stimuli consisted of 440 Hz square waves, bandpass filtered at 400 and 1500 Hz to attenuate onset transients, and played over headphones at a comfortable listening level.

Stimuli were presented, and responses collected, by a Turbo Pascal program (TWOTAP) with hardware-based millisecond timing resolution (Brysbaert et al., 1989). The apparatus for stimulus production was the same as that used in the rating and reproduction experiments. For response collection, TWOTAP read data from two MIDI drum pads (Roland MPD-4 and Dauz Designs), modified as described earlier, via a MIDI interface.

Design

A major concern in this experiment was the possibility that a subject might adopt a consistent tapping rate and maintain it over many or all trials. The risk would be highest if all stimuli were played at the same tempo, but even if several tempos were employed subjects might adopt several tapping rates, attending only to the tempo of each stimulus. In effect, this would reduce the experiment to a

tempo-identification task. To avoid this possibility, the number of tempos was not limited. Rather, each score was categorized as slow, moderate or fast according to its metronome marking (Table 4.2, left). The three tempos at which a given score was played (Table 4.2, right) were based on this categorization.

----- score tempo -----		----- stimulus tempo -----		
metronome	category	slow	moderate	fast
MM < 115	slow	MM	MM + 50%	MM + 100%
115 ≤ MM ≤ 165	moderate	MM - 33%	MM	MM + 33%
MM > 165	fast	MM - 50%	MM - 25%	MM

(MM = metronome marking in ♩s per minute)

Table 4.2 Tempo categorization of scores with metronome markings, and corresponding ranges of stimulus tempos.

The cut points used in categorizing scores (see Table 4.2) are the 33rd and 67th centiles of the distribution of metronome markings.

Almost all of the folk song scores included metronome markings. For the few that did not, and for the new rhythms, metronome markings were assigned randomly according to the shortest note value occurring in the rhythm, as shown in Table 4.3:

shortest note value	category	mm range	midpoint
♪	slow	65 - 115	90
♪	moderate	115 - 165	140
♩	fast	165 - 215	190

Table 4.3 Tempo categorization of scores without metronome markings.

In this experiment there were a total of 80 scores (57 folk songs and 23 new rhythms; see Appendix D), representing four time signatures x five error types x four scores. Each score was presented at each of three tempos, as described above, for a total of 240 stimuli. The design was run in three 40 - 60 minute sessions for each subject.

The 80 scores were arranged in three pseudo-random orders, A, B and C, and assigned tempos such that over the three orders each score was played once at each of its three tempos. No two consecutive scores had the same tempo. The three experimental sessions comprised orders A, B and C as well as their retrogrades, A', B' and C'. Four subjects were run in each of four orders: A B C, A' B' C', C B A, and C' B' A'.

Procedure

Subjects were asked to tap along with the rhythms using both hands, in such a way that a) each hand tapped isochronously and b) the two hands tapped together every second or every third tap. At the beginning of the subject's first session, he/she was given approximately 20 minutes of training and practice, which included tapping to a metronome at various tempos and tapping to sample rhythms, to ensure that the subject a) understood the nature of the tapping required and b) was able to perform such tapping. Only one subject had any apparent difficulty

with the tapping, and that subject was replaced.

For each session, the subject was seated in a sound-attenuating booth inside a small, closed room. The drum boxes were positioned approximately three feet apart. The subject rested his/her forearms on telephone books to bring them up to the level of the tops of the drum boxes. The subject initiated each trial by tapping either of the boxes; this was intended to encourage subjects to keep their hands on the boxes, ready to tap. Between the boxes was a small terminal, used to display various messages (to be described) to the subject. The terminal's cursor was turned off throughout the experiment so that its blinking would not interfere with the subject's task.

The session began with a series of practice trials designed to allow the subject to get used to the drum boxes. the stimuli and, finally, the task of listening and tapping. The practice session began with two isochronous sequences. For the first of these the subject was instructed (by a message on the terminal's screen) to tap two-against-one. For the second sequence the subject was instructed to tap three-against-one. The goal of these trials was to familiarize the subject with physical characteristics (e.g. gap, spring tension) of the drum boxes, with the motor and auditory feedback they provide, and with the task of synchronizing taps with the stimulus.

Next two musical rhythms, one with a duple meter and one with a triple meter, were presented, and again the subject was told how to tap. The goal of these trials was to familiarize the subject with the task of tapping synchronously to a non-isochronous stimulus and with the problem of finding the correct phase. Finally, two more scores, one with a duple meter and one with a triple meter, were presented, but now the subject had to determine how to tap. These trials introduced the subject to the full experimental task. Each practice stimulus was presented for as long as it took the subject to tap at least 25 times with each hand or for 60 seconds, whichever was shorter. All scores used as practice stimuli were scores that BEATS had analyzed correctly. A different set of four scores were used for each of the three experimental sessions. When they had finished the six practice trials, subjects were given the option of repeating the practice session. A number of subjects chose to do this in their first session, but few did in later sessions.

After the practice session(s), the subject heard the 80 experimental trials. Each trial began with the terminal asking the subject to tap either key when ready. When the subject tapped, the program began the stimulus after a random delay of 1500 to 2000 milliseconds. This delay was intended to prevent the subject from perceiving the 'ready'

tap as related to the rhythm. As the stimulus played, the subject began tapping as soon as he/she had determined the meter.

Except (as noted) in the practice trials, each stimulus was played only for as long as the subject took to tap at least ten times with each hand or for 60 seconds, whichever was shorter. This criterion was intended to standardize, as much as possible, the probability of motor error, both between subjects and between trials. The specific choice of ten taps was an attempt to occupy the middle ground between two undesirable outcomes. Transcripts of pilot subjects' responses showed that if more taps were required, some hesitant or slowly-tapping subjects would reach the end of the stimulus before reaching the quota, resulting in unequal numbers of taps recorded from trial to trial and from subject to subject. Worse, if fewer taps were required, some incautious subjects would use up their taps in exploratory tapping before settling on a particular metric interpretation, resulting in a needless inflation of the miscellaneous error category at the expense of specific error categories.

When the quota of taps was reached, the stimulus was stopped and a beep sounded from the terminal. The beep allowed subjects to distinguish the end of the stimulus from a pause in the stimulus due to a rest in the music.

Stopping the stimulus as soon as the quota of data was recorded also kept subject fatigue to a minimum and reduced the chance of the subject recognizing, and perhaps identifying, the stimulus.

Recognition Test

After a subject completed the three experimental sessions, he/she was given a recognition test. In this test the subject heard each of the 80 scores, played as in the experiment. To maximize the probability that if a subject knew a piece he/she would recognize it in this test, each score was played at its scored tempo. To minimize the chance that a subject would recognize in the test a piece he/she did not recognize during the tapping trials, each score was played for only as long as it was played on any tapping trial. Both of these restrictions incidentally helped to keep the test reasonably short (around 45 minutes).

Each recognition trial was initiated by the subject pressing a key on the terminal. After a brief random delay the test item began playing. When it was over the subject was asked if he/she had recognized the rhythm, and if so to name, sing, hum or otherwise identify the music from which the rhythm had been taken. These responses were recorded by the experimenter. There were several possible responses between a positive identification and an unambiguous 'no'.

If the subject merely reproduced the rhythm, without being able to indicate an associated melody, the trial was scored as a non-recognition. If the subject sang a melody that did not match that of the original music, its rhythm was notated, using standard musical notation, and compared with that of the stimulus. If these matched the trial was scored as a recognition. Correct tapping responses to stimuli correctly identified, by any method, were not used in further analysis. All other tapping responses were analyzed as described below.

Response Scoring

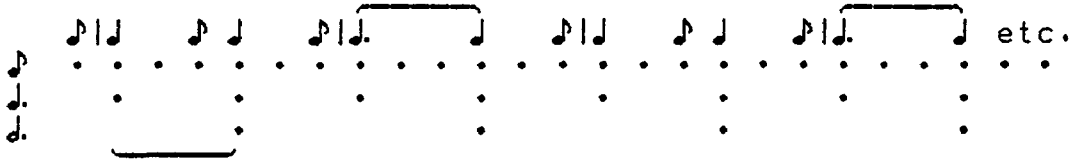
For each trial the computer produced a separate file containing an integrated record of the stimulus and the taps. These files were compared to the standard analyses by means of a computer program (SRGRAPH) which displays an integrated graphic image of the stimulus and response. These images are similar to the illustrations in this paper except that taps are shown in their exact temporal relation to the stimulus (see Appendix B for examples).

The basic principle behind the classification of tapping responses was that categories and criteria should facilitate, as much as possible, evaluation of the specific processes (e.g. Upbeat, Stretch) of which BEATS is made. For example, BEATS does not have any means of recognizing and correcting errors beyond the rules described. A

subject's first response therefore seems the most appropriate to compare with BEATS' analysis. With this principle in mind, the following classification scheme was devised.

First, for the purpose of classifying subjects' responses, each BEATS analysis is characterized by the following:

1. BEATS error type (correct, default, stretch, positive and negative phase)
2. The note values of the metric levels produced by BEATS.
3. For default errors, the note value of the level that represents the default error.
4. For stretch errors, the note value of the level created by the Stretch operation.
5. For phase errors, the note value of the lowest level at which the phase error is evident, and the note value representing the size of the (negative or positive) phase error. This is illustrated below (the rhythm is "The Farmer in the Dell"):



The phase error occurs at the ♩. level; the lower levels are in the proper relation to the score. The size of the phase error, as indicated by the bracket beneath the figure, is

the note value (♩.) corresponding to the interval between where a tick at the level in question occurs and where it should have occurred.

Second, a subject's tapping response is treated as a two-level metric hierarchy. The lower level, i.e. that with the shorter period, is referred to as level 1, and the higher level as level 2. Tapping responses were classified as described below. For each classification, the basic criteria are followed by additional criteria, if any, for specific BEATS classifications.

1. Correct response. Basic criteria: Both levels belong to a correct metric hierarchy and both have the correct phase.

Additional criteria:

when BEATS = default: Adding BEATS default level (to the tapping levels) results in an ill-formed hierarchy.
when BEATS = stretch: Adding BEATS stretch level results in an ill-formed hierarchy.

2. Correct/consistent response. Basic criteria: Both levels belong to a correct metric hierarchy and have the correct phase.

Additional criteria:

when BEATS = correct: Category not applicable
when BEATS = default: Both levels match BEATS' levels (but not the default level).
when BEATS = stretch: Both levels match BEATS' levels (but not the stretch level).
when BEATS = phase: Both levels are lower than BEATS' phase error level.

(This classification is necessary because when BEATS makes an error, some of the levels in the (incorrect) hierarchy may nevertheless belong to a correct hierarchy. If the subject taps at such levels, there is no way to know which hierarchy he/she had in mind. The correct/consistent classification reflects the fact that the response is correct but may belong to an incorrect hierarchy.)

3. Meter error. Basic criterion: One or both levels do(es) not belong to a correct metric hierarchy.

Additional criteria:

when BEATS = default:	One of the levels matches BEATS' default level; the other level may or may not belong to BEATS' hierarchy.
when BEATS = stretch:	One of the levels matches BEATS' stretch level; the other level may or may not belong to BEATS' hierarchy.
when BEATS = phase:	Level 2 has the same note value as BEATS' phase level but has correct phase, OR Level 2 does not have the same note value as BEATS' phase level.

4. Meter/different error. Basic criterion: Same as above but neither level matches BEATS' meter error level. Applies only to BEATS' default and stretch errors.

5. Phase error. Basic criteria: Both levels belong to a correct metric hierarchy and one or both do(es) not have correct phase, OR level 2 belongs to a correct metric hierarchy and does not have correct phase.

Additional criterion:

when BEATS = phase: One of the levels matches BEATS' phase level AND the size of the subject's phase error matches that of BEATS'.

6. Phase/different error. Basic criteria: Same as above but neither level matches BEATS' phase level OR the errors' sizes do not match. Applies only to BEATS' phase errors.

7. Miscellaneous error. Several types of responses, all of which defy classification as described above, fall in this category. In decreasing frequency order, they are:

a. The tapping interval is unstable.

b. There is no sustained tapping.

c. The tapping forms no hierarchy.

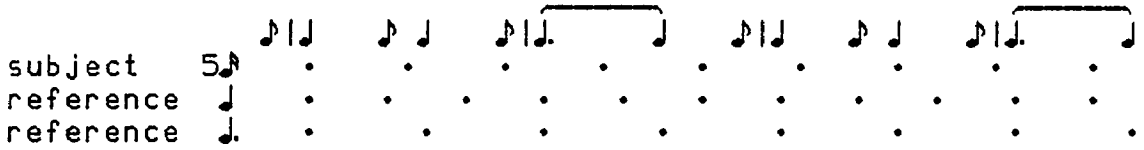
d. The subject did not tap.

In addition, policies were adopted to handle three types of anomalous response. In all cases the underlying principle is the fact that this experiment is more concerned with what meter the subject extracted from a given rhythm than with how well the subject executed the tapping task.

1. Despite instructions to the contrary, subjects occasionally tapped two levels in a 4:1 ratio (e.g. ♪ and ♩). In such cases the intermediate level (here, ♩) was

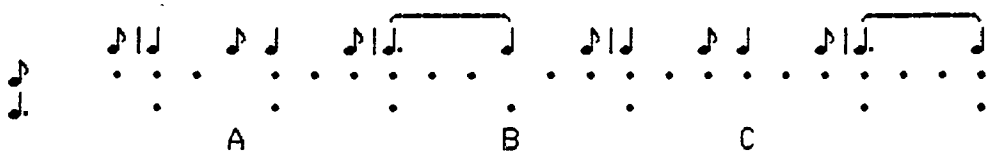
inferred for scoring purposes. This interpretation was supported by the fact that in many of these cases the subject switched from a 4:1 response to one of the two 2:1 responses implicit in it (here, ♪ and ♩ or ♩ and 𝅗𝅥).

2. Subjects occasionally adopted tapping intervals that did not unambiguously correspond to a note onset interval, or integral part thereof, that occurred in the stimulus. Here is an idealized example:



Here the subject's taps (one hand only) occur at intervals of five ♪s. Such an interval bears no obvious relation to any interval in the rhythm. It is assumed that this represents, at least in part, an error in execution rather than an 'error' of perception. On this account the subject extracted either a ♩ or 𝅗𝅥 interval but for some reason incorporated a uniform distortion in executing that interval. It being impossible to know which interval the subject extracted, such cases were classified as anomalous meter errors.

3. Sometimes a tap seems to be missing. Below are three examples:



Taps seem missing from level 1 at points A and B and from level 2 at point C. Because the context predicts each of these taps, it is simpler to suppose that the subject simply did not tap hard enough than to concoct an exotic error category. Such an interpretation is corroborated by the fact that this phenomenon is typically found in responses of subjects whose tapping is generally light.

All data were scored by the author, using SRGRAPH. All responses classified as agreeing with BEATS' predictions were independently scored by a disinterested third party trained in the use of SRGRAPH and the above classification scheme but ignorant of BEATS' predictions and of the purpose of the experiment. Examples of response classifications, using pilot data displayed in screen dumps from SRGRAPH, are in Appendix C.

Results and discussion

One subject recognized one stimulus, resulting in the removal of three trials from the analysis. The overall results are shown in Table 4.4 as column percentages. For example, when the stimulus was a score in which BEATS made a default error, 29.17% of subjects' responses were correct, 2.08% were correct but consistent with BEATS, 36.07% were meter errors that agreed with BEATS' error, 7.42% were meter errors that did not agree with BEATS'

error, 17.84% were phase errors, 6.12% were anomalous meter errors, and 1.30% were miscellaneous errors. An entry of na indicates that that cell's combination of stimulus classification and response classification (e.g. default and phase/different) cannot occur. The rightmost column pools all stimulus types and shows the percentage of subjects' responses that fell into each response classification, regardless of stimulus type.

Subject	BEATS correct	default	stretch	+phase	-phase	all stimuli
correct	54.95	29.17	23.44	29.04	13.15	29.95
correct/ consistent	na	2.08	4.30	12.76	11.07	6.04
meter	23.57	36.07	33.46	34.77	31.38	31.85
meter/ different	na	7.42	16.54	na	na	4.79
phase	13.67	17.84	17.84	11.85	35.55	19.35
phase/ different	na	na	na	6.38	5.99	2.47
anomalous meter	5.60	6.12	2.99	4.04	1.95	4.14
misc.	1.82	1.30	1.43	1.17	0.91	1.33

Table 4.4: Distribution of responses by stimulus type.

If BEATS and subjects are independent, the 'all stimuli' column should predict all of the other columns. That is, if stimulus classification has no bearing on subjects'

responses, the distribution of response classifications for each stimulus classification (first five columns) should be approximately the same as the distribution of response classifications over all stimulus classifications. Table 4.4 shows that this is not the case, and a test of independence shows that stimulus class and subjects' errors are indeed associated (chi square = 1090, $p < .001$).

To elucidate the nature of the relation it is necessary to make comparisons of cells within Table 4.4 and others. For example, to examine the effect of stimulus time signature, a table of subject error type by time signature was made for each stimulus classification; this amounts to subdividing each column in Table 4.4 into the four time signatures. The resulting table for correct stimuli is Table 4.5.

	time signature			
response	2/4	4/4	3/4	6/8
correct	120	101	103	98
meter error	29	65	35	52
phase error	24	21	37	23

Table 4.5 Distribution of responses by stimulus time signature for correct stimuli.

In such a table, effects of time signature appear as differences between two columns within rows. Tests of these differences were made by computing confidence intervals around the difference between two columns in a given row. For example, 2/4 and 4/4 time signatures can be

contrasted with respect to correct responses by dichotomizing the relevant columns in terms of correct/other responses, resulting in the fourfold table shown in Table 4.6.

response	time signature 2/4	4/4
correct	120	101
other	53	86
	173	187

Table 4.6 Contrast of 2/4 and 4/4 stimuli with respect to correct responses.

The contrast illustrated in Table 4.6 is a test of the null hypothesis that

$$p(\text{correct} \mid 2/4) = p(\text{correct} \mid 4/4)$$

The difference between these probabilities is 0.15. The standard error of estimate is the product of the square root of the pooled binomial variances for the two columns and the square root of the critical chi square value for (columns - 1) degrees of freedom (Marascuilo, 1971). In this example the 95% confidence interval around the difference between the two columns is 0.15 +/- 0.14; because zero falls outside of this interval the null hypothesis can be rejected, indicating that correct responses are significantly more likely for 2/4 than for 4/4 stimuli. The analysis is completed by computing similar contrasts for each of the six pairs of columns in

Table 4.5 at each of the three rows. This post-hoc method makes it possible to find the differences within a table that led to rejection of the null hypothesis of independence. Rather than presenting a raft of statistics, what follows is a summary of the results of the post-hoc tests relevant to each of a number of questions about the data.

Overall Patterns. Correct responses are more likely for correct stimuli than for any others (as expected), and less likely for negative phase errors than for any others. Meter errors are less likely for correct stimuli than for any others, but equally likely for default, stretch, and positive and negative phase errors. Phase errors are more likely for negative phase errors than for all other stimuli, and more likely for BEATS' meter errors than for positive phase errors. These patterns, however, do not take correct/consistent responses into account. Correct responses are equally likely at all tempos. If correct/consistent responses were in fact correct, they should likewise be equally distributed across tempos. However, such responses are more likely at slower tempos. This suggests the following explanation. First, the subject has constructed the same metric hierarchy as BEATS. Second, the meter or phase error level is generally not among the lowest two or three levels of the hierarchy.

Therefore, if the subject taps levels that are low in the hierarchy, the response is unlikely to reflect BEATS' error and will be classified as correct/consistent. Handel (1984) has demonstrated that subjects tap lower metric levels at slower tempos, and the same relation appears in this experiment (see below). This accounts for the greater likelihood of correct/consistent responses at slower tempos, and suggests that correct/consistent responses would have been meter or phase errors, rather than correct responses, if the subject had tapped at higher levels of his/her (incorrect) hierarchy. Accordingly, correct/consistent cells were removed. For default and stretch stimuli their frequencies were added to meter errors, and for both phase error types they were added to phase errors. The adjusted overall results are shown in Table 4.7.

Subject	BEATS correct	default	stretch	+phase	-phase	all stimuli
correct	54.95	29.17	23.44	29.04	13.15	29.95
meter	23.57	38.15	37.76	34.77	31.38	33.12
meter/ different	na	7.42	16.54	na	na	4.79
phase	13.67	17.84	17.84	24.61	46.62	24.11
phase/ different	na	na	na	6.38	5.99	2.47
anomalous meter	5.60	6.12	2.99	4.04	1.95	4.14
misc.	1.82	1.30	1.43	1.17	0.91	1.33

Table 4.7 Distribution of responses by stimulus type, adjusted for correct/consistent responses.

Phase errors are now more likely for both positive and negative (BEATS) phase errors than for all other stimuli, but the difference between positive and negative is still significant. Two fifths of the stimuli predict meter errors, and another two fifths predict phase errors. The overall proportions of meter and phase errors are now closer to, but still short of, these proportions. Much of the difference is accounted for by meter/different and phase/different responses. However, there is no valid reason to include such responses with meter or phase errors, so they must remain separate. The remaining two error types (anomalous and miscellaneous) account for 5.5%

of responses. Anomalous meter errors are more likely for correct and default stimuli than for negative phase errors, but it is not clear why this should be so. It happens that 80.5% of these errors were made by five subjects; conceivably the unequal distribution of anomalous meter errors across stimulus types reflects idiosyncracies of these subjects. Finally, miscellaneous errors are equally likely in all stimulus categories. Over all, BEATS correctly predicts 40.4% of the responses.

The Effects of Tempo. The differences described above are not duplicated in every particular when the data are broken down by tempo, but the overall pattern persists. At all tempos, correct responses are more likely for correct stimuli than for all others, and less likely for negative phase errors than for all other stimuli. Phase errors are more likely for negative phase stimuli than for all others. BEATS does not consider tempo and therefore predicts no differences due to tempo. This was generally borne out, but there was an interesting exception. Phase errors are more likely for positive phase stimuli at slow tempos than at fast tempos, and the reverse is true for correct responses to the same stimuli. Phase errors for negative phase stimuli also decline slightly, though not significantly, with increasing tempo. In both cases the effect may reflect the fact that at a faster tempo more

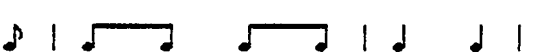
notes may be available, in sensory registers and in short term memory, thus increasing the likelihood that the listener will be able to disambiguate the phase of the stimulus. For example,

might be ambiguous; it can be heard as

or as

If more notes are available,

the ambiguity is resolved:



This effect might be simulated in BEATS by means of an 'input window' (Jones, Miller & Scarborough, 1988), a control structure that determines how much of the input (the score) is available for processing at a given moment. Such a structure is in some respects a generalization of the retrospection parameter in BEATS (see p. 24). If the window corresponds to a span of time (rather than, say, a number of notes), it should be possible to duplicate the effect described here.

It was mentioned earlier that subjects tap at higher metric levels (i.e. tap intervals correspond to longer note durations) at faster tempos. This prediction, based on

work with polyrhythms by Handel and his colleagues (Oshinsky & Handel, 1978; Handel & Oshinsky, 1981; Handel & Lawson, 1983; Handel, 1984; also Beauvillain, 1983), was borne out in the present experiment. Figures 4.1 and 4.2 show subjects' choices of tapping levels as a function of tempo for each of the four stimulus time signatures. For 2/4 stimuli, for example, subjects preferred to tap the ♪ and ♩ levels at the slowest tempos, while at tempos above approximately 100 ♩/minute they preferred to tap the ♩ and 𝅗𝅥 levels. This preference declined at tempos above approximately 200 ♩/minute, where subjects chose to tap the 𝅗𝅥 and 𝅝 levels equally often. Analogous patterns of preference obtain for the other time signatures. In all cases, subjects tended to tap at lower metric levels at slower tempos and at higher metric levels at faster tempos. This pattern is likely a manifestation of Lerdahl & Jackendoff's (1983) concept of the tactus, an intermediate rate of pulses that is most salient in a listener's perception of a piece of music. Because the tactus is, at least to some degree, independent of tempo, it will correspond to a low metric level at slow tempos and to a higher metric level at faster tempos, just as subjects' taps do in this experiment and in Handel's studies. The large and non-systematic departures from this general pattern are due to the fact that a given tempo/time

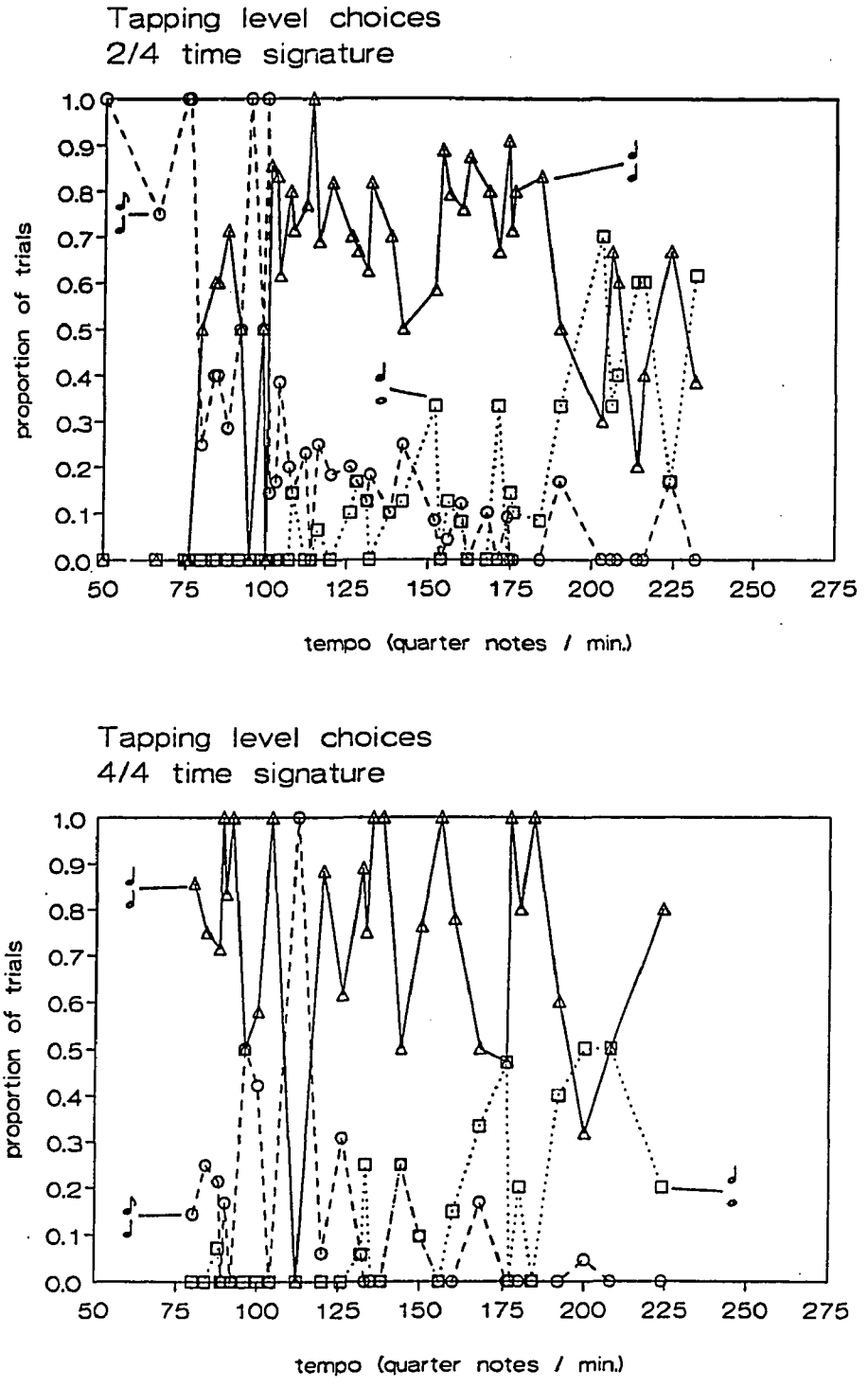


Fig. 4.1 Tapping levels as a function of tempo for 2/4 and 4/4 stimuli in the error matching experiment.

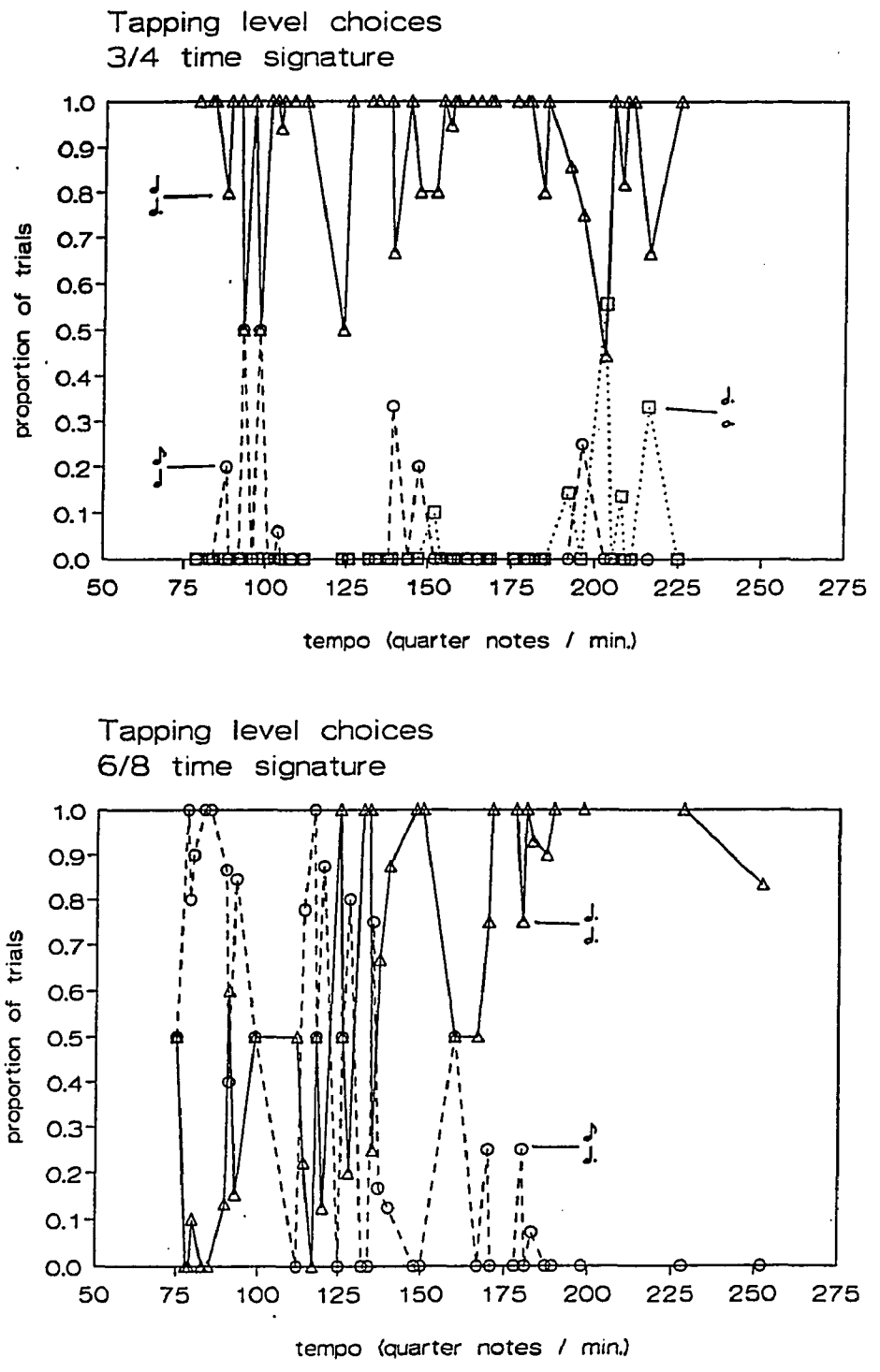


Fig. 4.2 Tapping levels as a function of tempo for 3/4 and 6/8 stimuli in the error matching experiment.

signature combination represents only one or two particular stimuli. Presumably there are characteristics of a rhythm other than tempo, such as the proportions of short or long notes, that are also relevant to a subject's choice of tapping levels. These data provide a replication of Handel's polyrhythm results in the context of more conventional rhythms, and also support Lerdahl & Jackendoff's ideas about the tactus.

The Effects of Time Signature. Here again the overall pattern is found for each time signature. However, a time signature, unlike a tempo, reflects certain aspects of a score's metric structure, so it is not surprising that there are some specific differences in the distributions of response types across time signatures, as summarized in Table 4.8 (percentages are by column; they do not sum to 100 because of the response types not listed).

	time signature			
response	2/4	4/4	3/4	6/8
correct	41.6	28.6	25.3	24.3
meter	19.7	34.6	43.0	35.2
phase	24.2	23.6	19.3	29.4

Table 4.8 Distribution of responses by stimulus time signature, expressed as column percentages.

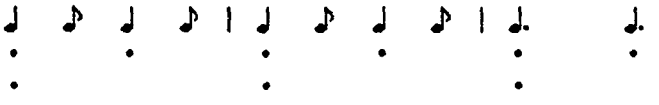
In general, correct responses are more likely for 2/4 stimuli and meter errors are less likely. This may reflect the fact that a measure in 2/4 is shorter, measured in

terms of the number of notes of a given value it contains, than a measure in any of the other time signatures. For example, there are four ♪s in a 2/4 measure, six in a 3/4 or 6/8 measure, and eight in a 4/4 measure. While ties across bar lines are relatively uncommon in folk songs, the organization of relative durations within a measure is comparatively unrestricted. This can be illustrated by the following simple rhythm:

The shortest periodicity here is 𝅗𝅥, the duration of the 2/4 measure. If the measure tends to be the shortest periodicity, it follows that in a given span of time, the listener hears more equal-duration note groups in a 2/4 rhythm than in other time signatures. This in turn increases the likelihood that the listener will discover an appropriate periodicity to predict future events. If this proves to be a reliable effect, the input window described earlier may provide a way to produce the same effect in BEATS.

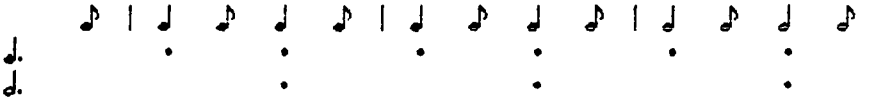
Meter errors are in general more likely for 3/4 stimuli than for other time signatures; this is true for default and positive and negative phase errors, but not for stretch errors. This may reflect the default status of duple meters; just as BEATS must take extra steps to extract a triple meter, so may listeners. This argument is

consistent with the fact that meter errors are more likely for 3/4 default stimuli than for 3/4 stretch stimuli. If duple meters are a default value, default 3/4 stimuli will be incorrectly perceived if this default is allowed to operate. This would seem to predict a similarly high rate of meter errors for 6/8 stimuli, but in fact subjects often tapped duple meters to such stimuli:



BEATS uses the Stretch operation to establish ♩ ♪ as a single metric unit. By doing the same, subjects can treat 6/8 rhythms in a duple manner.

Phase errors, for both positive and negative phase stimuli, are more likely for 6/8 rhythms. Virtually all of these errors are of the sort shown below:



Here the ♩. metric units cross bar lines, so the response is considered a phase error. However, the rhythm contains no clue to the proper phase; it is perfectly ambiguous. Put differently, removing pitches, accents and lyrics from a piece of music has different effects for different time signatures. The evidence described here suggests that 6/8 rhythms depend more heavily on these dimensions for cues to their meter. By contrast, 2/4 rhythms seem to have a

higher degree of redundancy between note durations and the other dimensions.

The Effects of Session and Stimulus Order. Subjects' agreement with BEATS' predictions improved over the course of the three sessions in this experiment. There were no significant differences across sessions in correct responses or in either of the error categories, but there was a significant decrease in both anomalous meter errors and miscellaneous errors. The improvement, then, was not in meter extraction but in tapping performance. The decline in anomalous meter errors over sessions supports the hypothesis, offered earlier, that these are errors not of perception but of execution.

There were three stimulus orders plus their retrogrades. A given subject's three sessions consisted either of normal or retrograde orders. The only effect of stimulus order was a smaller number of correct responses and a larger number of anomalous meter errors in the normal orders than in the retrograde orders. Closer inspection shows this effect to be purely coincidental. As mentioned earlier, five subjects were responsible for more than 80% of the anomalous meter errors. As it happens, four of these subjects were run in normal orders. These subjects also made relatively few correct responses.

Table 4.9 compares the percentage of responses in a

given response category across subjects. As shown, subjects differed most with respect to correct responses. Analysis of variance for each response category compared the 16 subjects with respect to the mean number of responses in that category across the three tempos. Subjects differed significantly with respect to correct responses and meter errors.

% of responses	mean	s.d.	$F_{15,32}$	p
correct	29.9	10.9	9.6	< .01
meter	31.7	5.5	2.4	< .05
phase	19.4	4.5	2.0	> .05

Table 4.9 Mean proportions of responses by error type across subjects.

Subjects who made more correct responses also tended to agree with BEATS more often. Over the 16 subjects there is a correlation of 0.67 between number of correct responses and number of responses that agree with BEATS. This may in part reflect the fact, mentioned earlier, that subjects who made fewer correct responses made more anomalous meter errors (and hence agreed with BEATS less often). Overall there is a correlation of -0.72 between correct response rates and anomalous meter/miscellaneous error rates. By contrast, anomalous meter/miscellaneous error rates have a slight positive correlation with meter errors ($r = 0.4$) and only a weak negative correlation with phase errors ($r = -0.26$). This suggests that anomalous meter errors and

miscellaneous errors were somewhat more likely to be made at the expense of correct responses than other response categories.

While there were no effects of either session or tempo, there remains the question of to what extent each subject's responses to a given score agree with one another. This is reflected in the number of different response categories a subject's responses fall into for a given score, as shown in Table 4.10. Because subjects heard each score three times, the number of different responses to a given score ranges from one, reflecting complete consistency, to three, reflecting no consistency.

		no. of different responses	
		mean	s.d.
stimulus type:	correct	1.99	0.76
	default	2.11	0.75
	stretch	2.19	0.69
	+phase	2.18	0.76
	-phase	2.12	0.74
	all stimuli	2.12	0.74

Table 4.10 Mean number of different responses over three trials.

Note that consistency is reckoned in terms of response classification rather than in terms of which levels were tapped. The reason for this choice is that different response classifications reflect different metric hierarchies, while different levels reflect tempo and other, idiosyncratic, factors. Subjects were slightly more consistent with correct stimuli, but what is more striking

is that on the whole subjects were not terribly consistent, both within and across subjects. BEATS, on the other hand, is nothing if not consistent. This difference points to some interesting conclusions.

BEATS is entirely deterministic. Listening to a given piece, it begins its analysis at the same point every time; specific events always have the same salience, and the same operations produce the same effects. BEATS is not a hypothesis that human perception is similarly deterministic, but rather a hypothesis that non-determinate aspects of perception are small enough that a deterministic model like BEATS will provide a reasonable approximation. Tempo is the only factor that varies over a subject's three trials with a given rhythm, yet there is no tempo effect. The fact that responses vary over repetitions without reflecting the effect of any experimental factor (i.e. tempo) suggests that the non-determinate aspects are substantial. The consistency problem shows that while over the long run BEATS' analyses have some predictive value, over the short run they have much less.

On the other hand, the apparent indeterminacy may be an artifact of requiring subjects to extract a meter from fewer cues than they are accustomed to. As demonstrated by Experiments 1 - 3, note durations alone do often carry sufficient cues to a piece's meter; BEATS' analysis matches

the standard about 40% of the time, listeners' analyses about 30%. But the other side of the coin is that more often note durations do not lead BEATS or listeners to the correct meter. This must mean that rather than being entirely redundant with duration, pitch and accent information are often necessary to complete or disambiguate the metric cues provided by duration. What is needed is knowledge of the interaction of musical dimensions during the extraction of meter. Duration-only stimuli may not be the appropriate way to determine the limits of duration-based meter perception, because if a subject agrees with BEATS on a given trial he/she is fairly likely to disagree on the next trial. Every rhythm yielded at least one correct response from one subject. BEATS' prediction should perhaps be thought of as a central tendency of a subject's response in the absence of all the expected cues.

A better approach might be to use stimuli that combine duration and pitch. By constructing stimuli in which melodic and temporal metric cues either agree or disagree with one another, it should be possible to explore the conditions in which particular types of cues from the two dimensions are used or not used, as the case may be. For example, in a duration-only sequence beginning with a short note followed by a longer note, the short note is often treated as an upbeat. By adding various consonant and

dissonant melodic cues to such a stimulus it can be determined whether some melodic cues increase the likelihood of this interpretation and others diminish it. If this is the case, then it could be argued that a rhythm without any melodic cue(s) at all is ambiguous and will give rise to different metric interpretations from trial to trial. Exploring the degree to which melodic cues do or do not override duration cues (and vice versa) may offer a more detailed characterization of duration as a cue to meter.

In summary, this experiment has replicated, in a broader context, one of the principal findings of the probe experiments, namely that listeners can and do find meters on the basis of durational cues alone. It has also demonstrated that duration is by no means the only cue to meter, although it is sometimes sufficient. From this perspective, trying to improve BEATS' performance within the current duration-only context would be mostly futile. Rather, the next stage in the development of BEATS should be the integration of melodic (and perhaps other) information in the stimuli, for example by implementing Lerdahl & Jackendoff's (1983) grouping rules (Jones, Miller & Scarborough, 1990), and by adding a set of heuristics to guide the interaction between dimensions.

4. General Discussion

There are several general conclusions to be drawn from these experiments. All the experiments support the idea that meter is represented hierarchically. In the probe experiments this support comes from the fact that all dependent variables reflect, to some degree, the metric level of each probe position as identified by BEATS. In the error-matching experiment, the agreement between subjects and BEATS indicates that BEATS correctly identifies metric levels. While this does not necessarily support the specific assumptions and processes of BEATS, it does indicate that a metric hierarchy is an appropriate goal of any model of meter perception.

The present results may seem, at first glance, to conflict with the experiments of Palmer and Krumhansl (1987 a,b), which suggested that meter and melody have separate and independent mental representations. Specifically, Palmer and Krumhansl found that subjects' ratings of the goodness of a musical phrase were very well accounted for by a linear combination of ratings of two incomplete versions of the same phrase: in one all notes had the same pitch, and in the other all notes had the same duration. Experiment 4, by contrast, points strongly to an interaction of meter with other musical dimensions. However, there is an important difference between the two

experiments. Palmer and Krumhansl's results concern representation of meter (and melody), while Experiment 4 points to interaction of pitch and duration in perception of meter. Independent representations of meter and melody, as such, are not necessarily inconsistent with the idea of pitch as a cue to meter or duration as a cue to melody. There is no reason to assume that perceptual interaction and representational independence are mutually exclusive.

With respect to BEATS itself, and meter-perception models in general, the results reported here present something of a paradox. The probe experiments show that listeners can extract metric structure from short sequences consisting only of durations, suggesting that the duration-only approach and efficient algorithms of BEATS constitute a plausible model of a demonstrated perceptual ability. However, when BEATS and subjects are compared in the context of extended duration sequences, a different picture emerges. The fact that BEATS and subjects often disagree is itself not surprising; it was never assumed that BEATS would be the entire story or that duration is the only cue to meter. What is perplexing is the fact that individual subjects tended to produce more than a single metric interpretation of individual stimuli. This inconsistency constitutes a deeper disparity between subjects and BEATS than the specific disagreements between BEATS' analyses and

subjects' tapping responses.

The resolution suggested here can be summarized as follows. BEATS may, after all, be a reasonable model of duration-based meter perception, with two qualifications. First, listeners normally use duration cues in conjunction with other (i.e. melodic, harmonic, dynamic) cues. Second, when these other cues are missing, duration cues are often inadequate or ambiguous. Why listeners cannot consistently make efficient use of duration cues, as BEATS does, is not clear. It seems likely that the answer to this question lies in a model of the interaction of all available cues in the perception of meter. Future research will be concerned with developing such a model.

A final question concerns BEATS' generality with respect to different musical cultures. In its present form, BEATS successfully analyzes rhythms of an idiom that might loosely be called Western. Why can't it analyze rhythms from other idioms? The answer involves two different questions. First, <u>should</u> BEATS be expected to work for more than one musical idiom? It can be argued that human listeners, like BEATS, understand only their native idiom. It is therefore appropriate for BEATS' rules to exploit the temporal characteristics of the Western rhythmic idiom, rather than using algorithms that attempt to discover temporal patterns in the music (e.g. Simon,

1968), processes that are as likely to succeed in one idiom as in another. Second, could BEATS succeed in a different idiom? Just as the musical knowledge embodied in BEATS' rules restricts BEATS to the idiom from which that knowledge is drawn, replacing or modifying some of the rules would allow BEATS to operate in a different musical domain. Different rules could be used without any changes to the basic machinery that moves through a score and maintains the time frame. However, only one set of rules can be used in any one BEATS implementation. This implies that a listener who is comfortable in more than one idiom has more than one set of rules, rather than a single more powerful or more general set of rules. Such listeners are to music as bilinguals are to language, operating with one set of rules at a time without interference from the other, yet able to move easily from one set to the other as the need arises. These implications of the BEATS approach suggest that investigations of listeners' abilities in multiple idioms may offer additional means of evaluating BEATS and other models like it.

Appendix A: BEATS' Rules

Wherever relevant, BEATS' current position is indicated by an asterisk. An asterisk at a note symbol indicates that the current position is the note's onset.

Setup

IF 1)  the onset of the second note has been reached

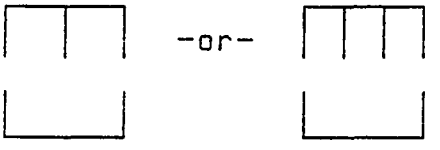
and 2) the time frame has not been created

THEN  create the time frame.

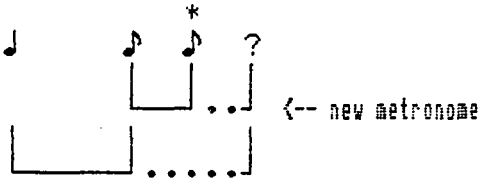
Induce

IF 1) 

a new onset interval is found

and 2)  -or- etc.

the interval stands in an integral ratio to the nearest existing metronome

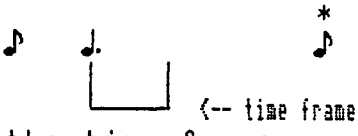
THEN 

add a new metronome corresponding to the new onset interval.

Longnote

IF 1) the Upbeat rule has just applied

and 2)



the time frame predicts an onset earlier than one actually occurs

THEN

enlarge the time frame so that it predicts the next onset, and create a new metronome.

Double

IF 1)

the time frame predicts a future note onset

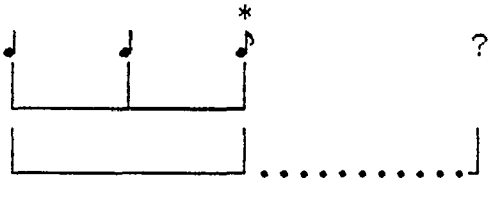
and 2)

the hypothesis is confirmed

and 3)

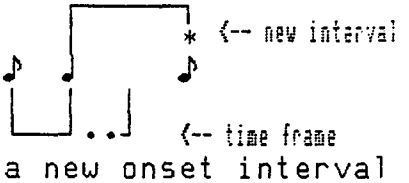
the time frame is smaller than some limit (e.g. o)

THEN



double the time frame and make a new metronome.

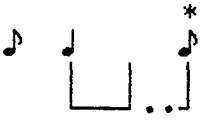
Upbeat



IF

a new onset interval is longer than any previous interval and longer than the time frame

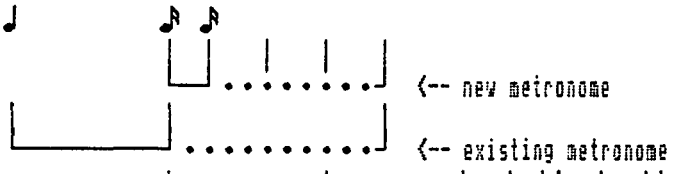
THEN



the first onset of the new interval becomes the downbeat.

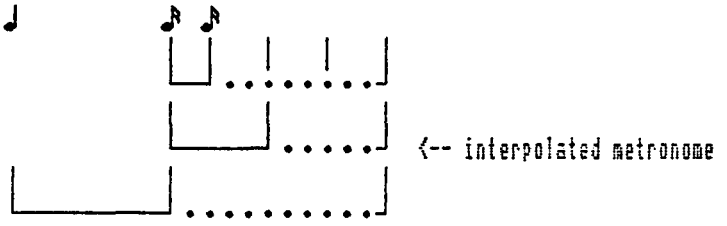
Interpolate

IF



a new metronome is created that divides the nearest existing metronome by >3 or <1/3

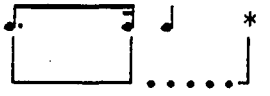
THEN



interpolate a third metronome that stands in a 1:2 or 1:3 ratio with both of the first two metronomes.

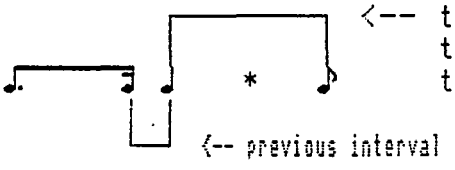
Stretch

IF 1)

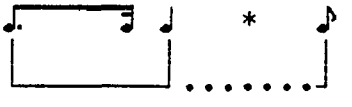


an onset occurs before the next predicted onset

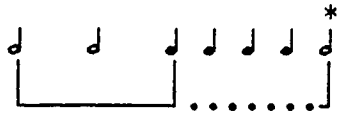
and 2)



THEN

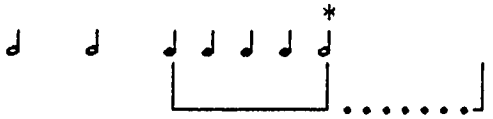


Slide

IF 1) 

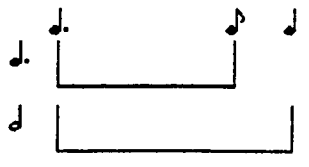
the onset predicted by the time frame is reached

and 2) a Double is not possible because the size limit has been reached.

THEN 

Slide the time frame by moving T1 to T2, T2 to T3, and T3 to T2 + (T2 - T1).

Remove (case 1)

IF 

the ratio between two adjacent metronomes is non-integral

and no interpolation is possible

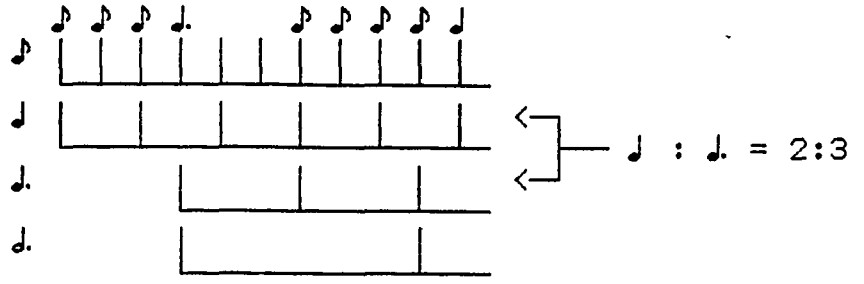
and ♩. : ♩ = 3:4
one of the metronomes is not consistent with the largest metronome

THEN 

remove the inconsistent metronome.

Remove (case 2)

IF



the ratio between two adjacent metronomes is non-integral

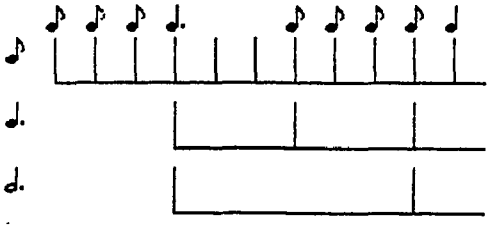
and no interpolation is possible

and ♩ : ♩. = 1:3 ♩. : ♩. = 1:2
each of the metronomes is consistent with the largest metronome

THEN ♩. / ♫ = 3 ♩. / ♬ = 2
estimate the meter by finding the larger of 2x or 3x the smallest metronome that integrally divides the largest metronome (in this case, 3)

and ♩ / 3 = ? ♩. / 3 = ♪
determine which of the problem metronomes is inconsistent with the meter (in this case, ♩)

and



remove the inconsistent metronome.

The order of priority of the above rules is

1. Setup
2. Induce
3. Longnote
4. Upbeat
5. Stretch
6. Double
7. Remove
8. Interpolate
9. Slide

Appendix B: Rating Experiment Instructions

In this experiment you will hear 126 short rhythms. Each rhythm will be four or five notes long. As each rhythm is played, listen carefully. When the rhythm is over, you will be asked to judge how well the last note of the rhythm fits the rest of the rhythm. You will make your rating using a scale of 1 to 7, where a rating of 1 indicates a poor fit, and a rating of 7 indicates a good fit.

1 2 3 4 5 6 7

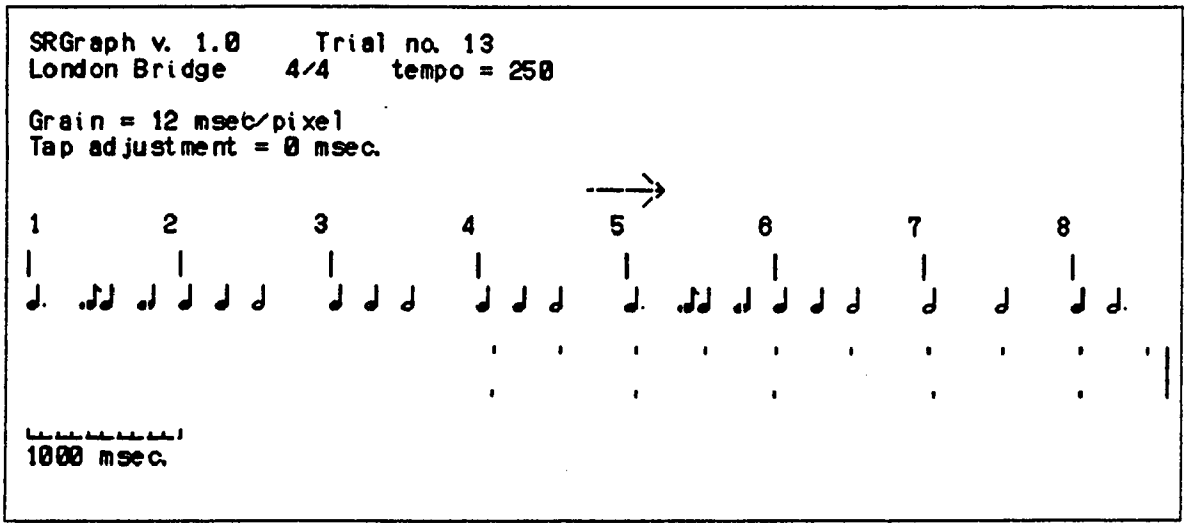
poor fit <----------> good fit

This rating scale will be displayed on a card in front of you throughout the experiment. In making your judgments, try to use the entire scale. Rhythms with the worst fit should be rated '1' and rhythms with the best fit should be rated '7'. (You will hear some examples before the experiment begins.)

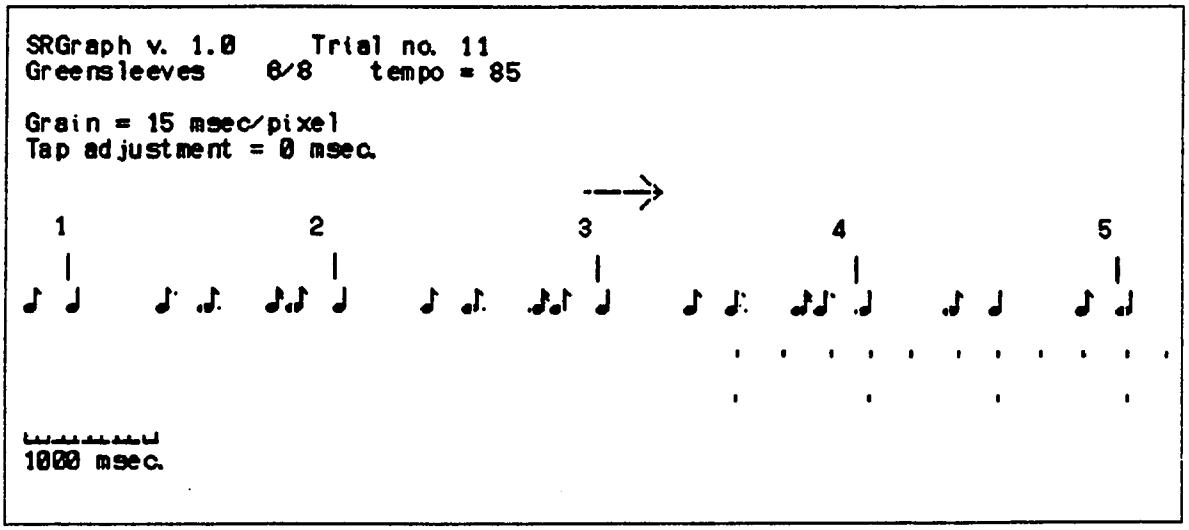
[The remainder of the instructions concerned using the terminal, the presentation of the examples, and the practice session.]

Appendix C: Tapping Response Examples

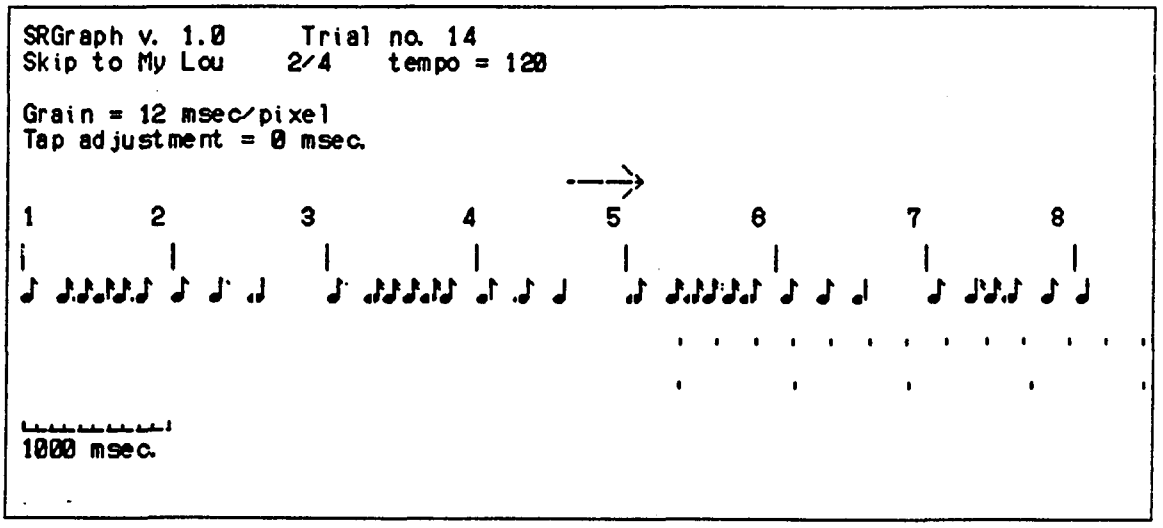
1. Correct response - duple meter (entire response shown).



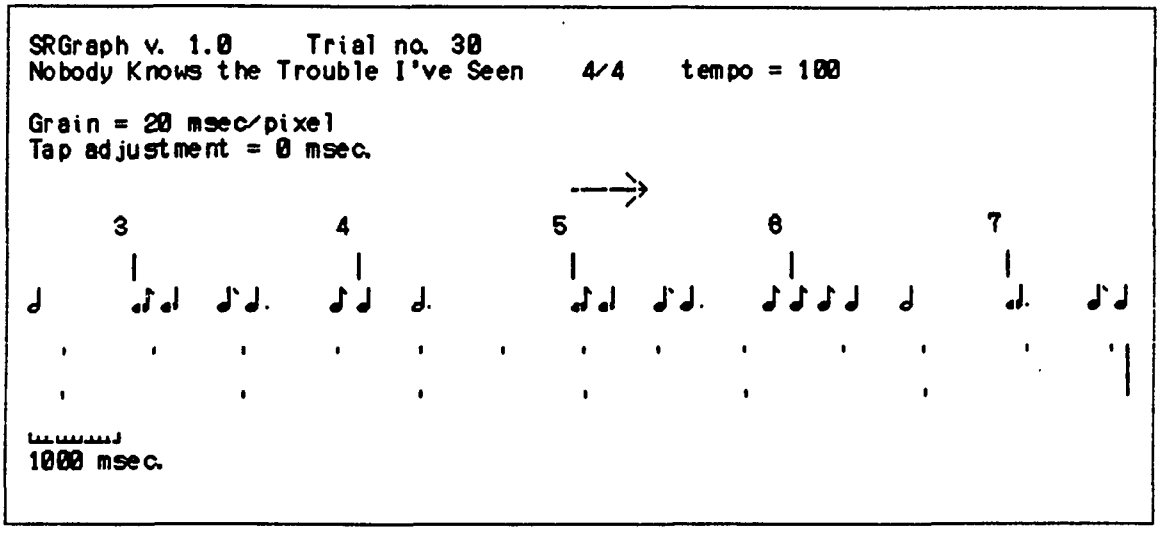
2. Correct response - triple meter.



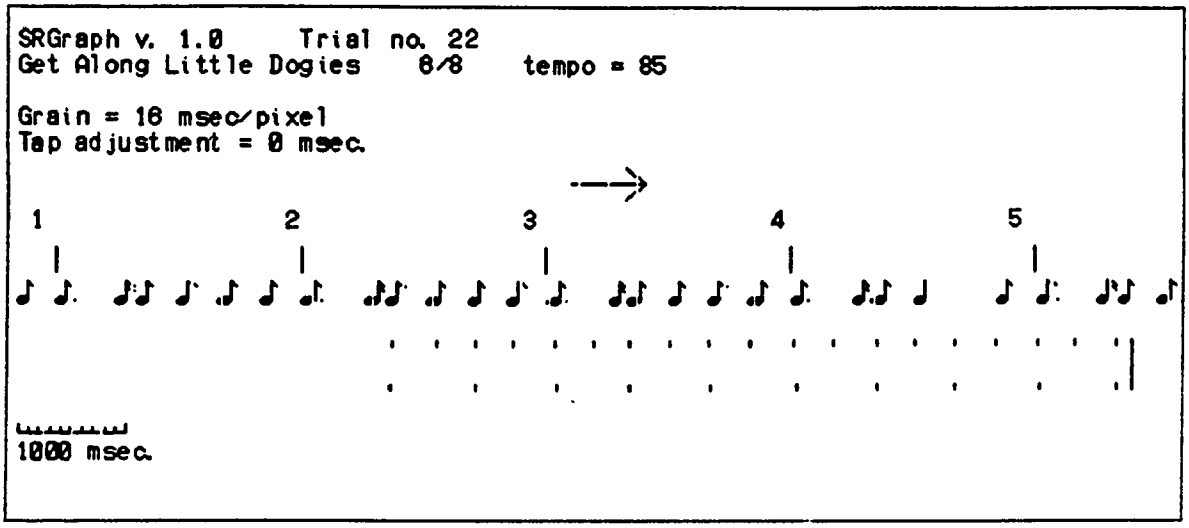
3. Meter error - duple meter (entire response shown).



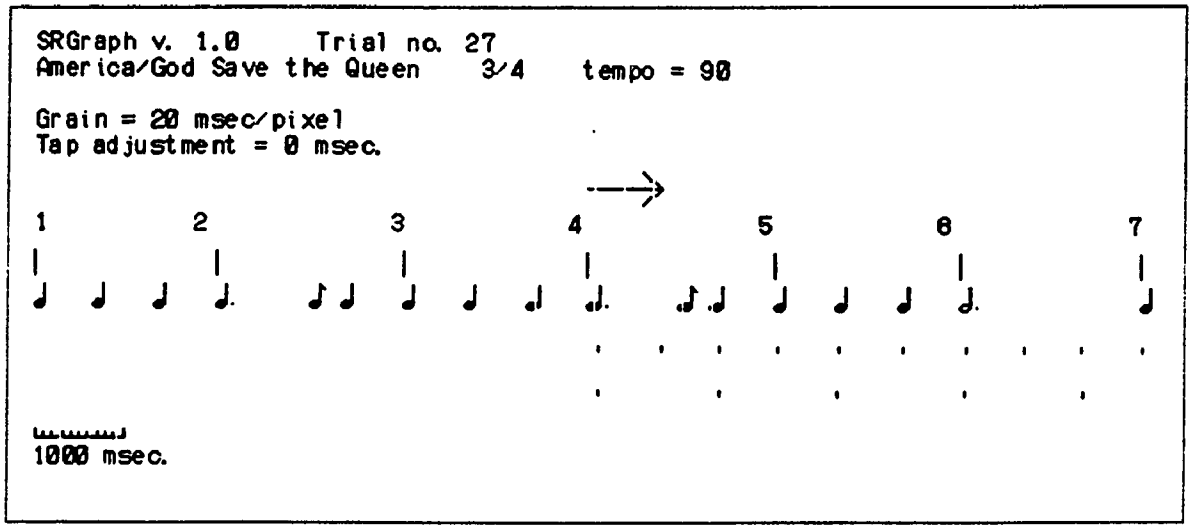
4. Meter error - duple meter (entire response shown).
Here the tapping is duple but the units (♩. and ♩.) are wrong.



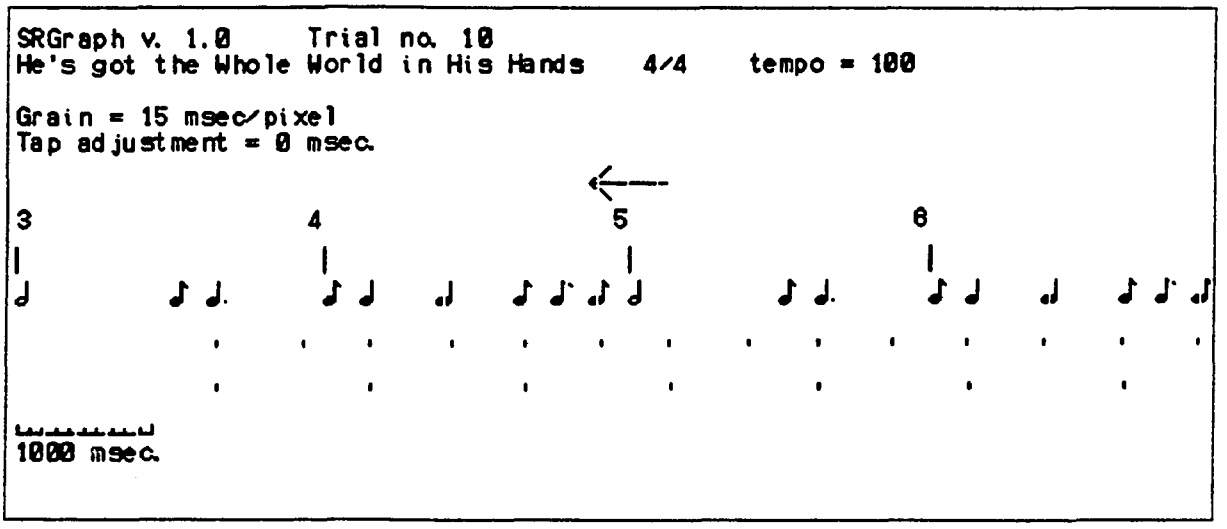
5. Meter error - triple meter (entire response shown). Tapping is reasonable but doesn't agree with the standard analysis.



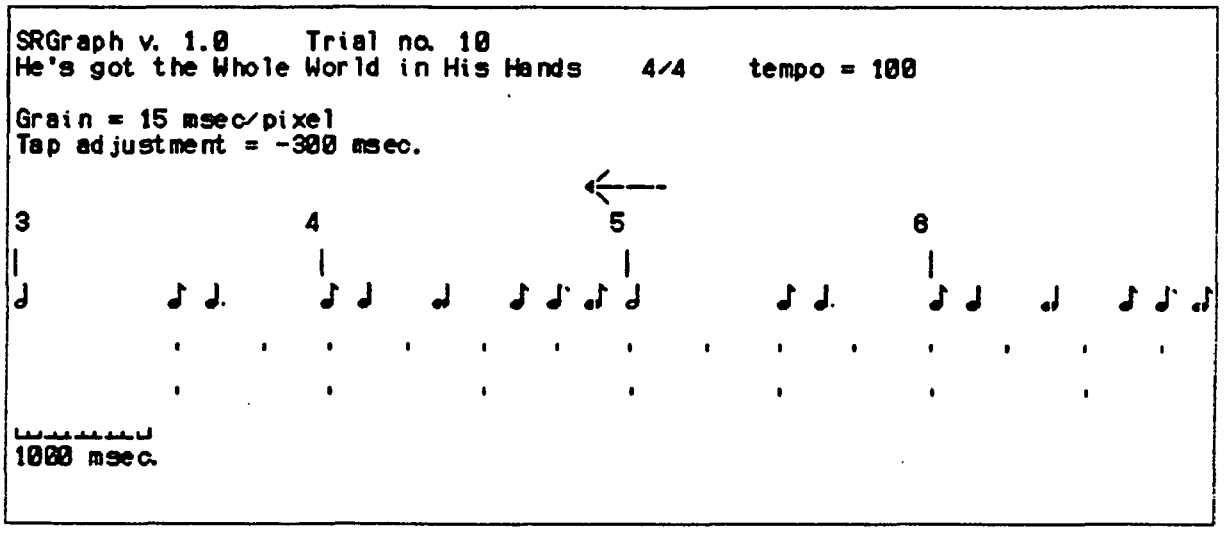
6. Meter error - triple meter.



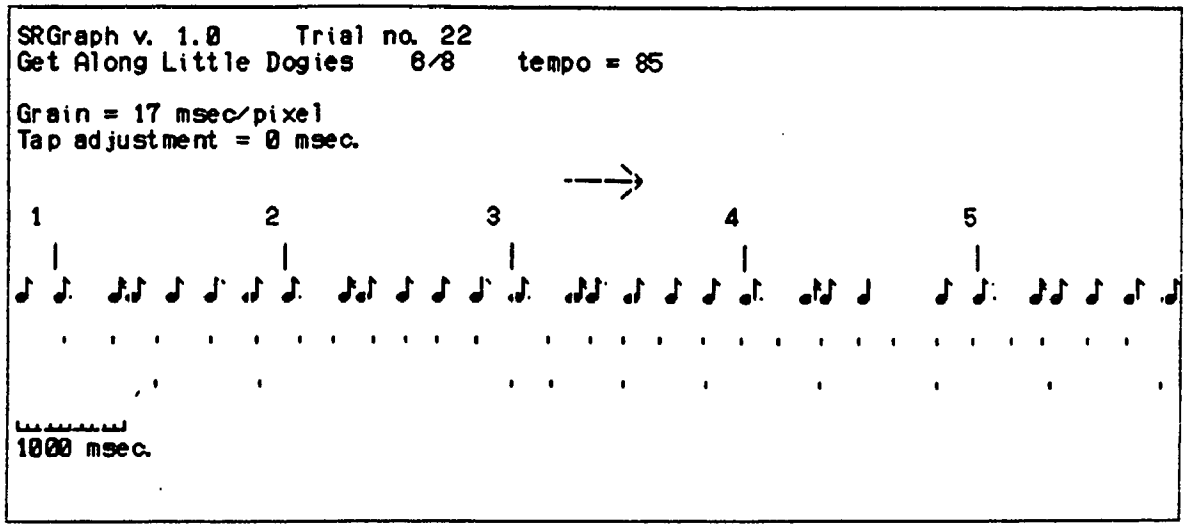
7a. Phase error - duple meter.



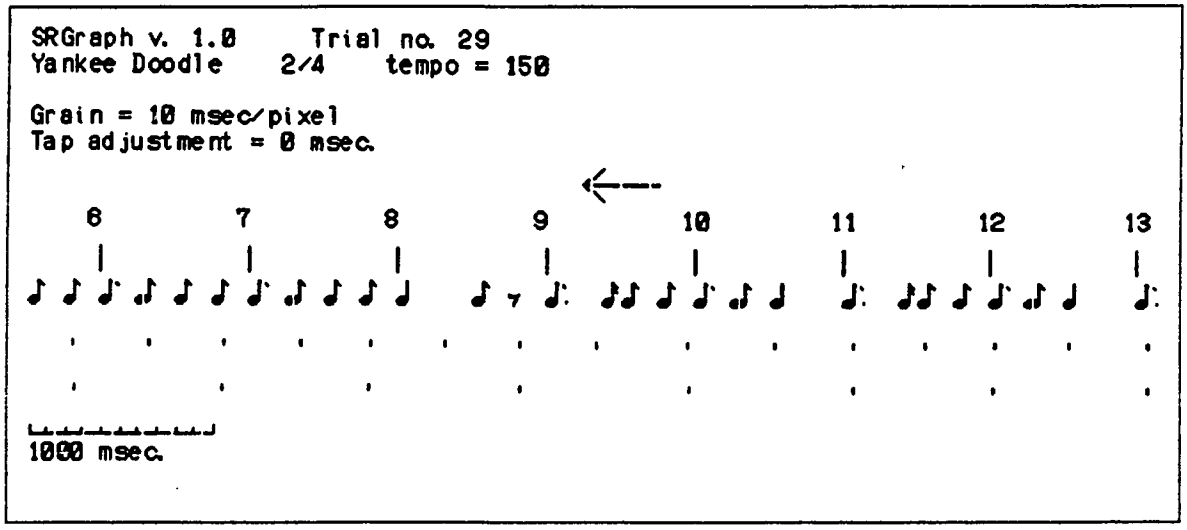
7b. Shifting the taps 300 ms. to the left shows that the above meter is correct; only the phase is wrong.



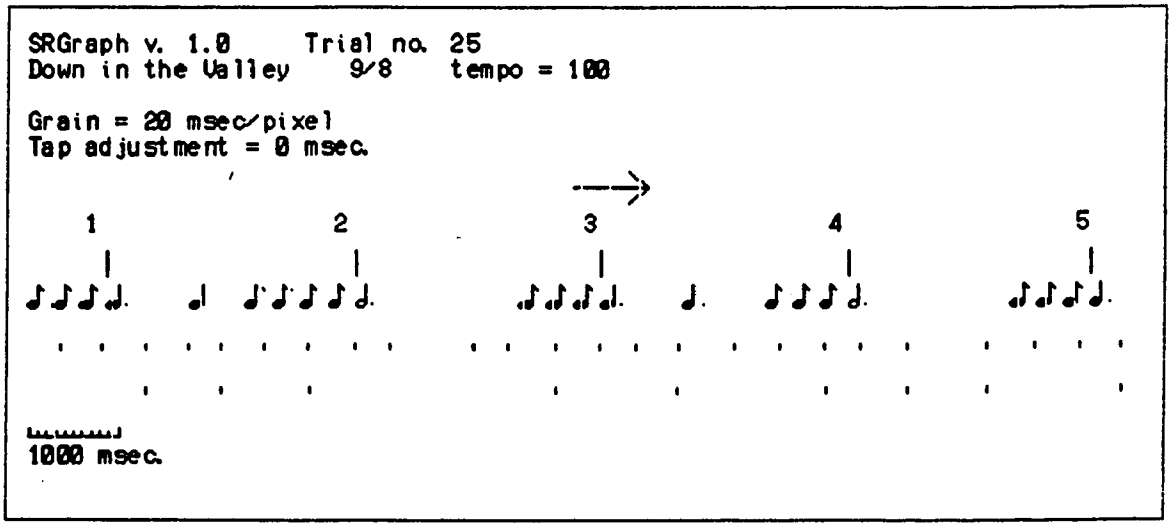
8. Phase error - triple meter.
An ambiguous beginning resolves into a phase error.



9. Phase error - duple meter. Subject corrects phase at measure 10, but first response is clearly a phase error.



10. Miscellaneous error - there is no stable response.



Appendix D: Stimuli for Experiment 4

2/4 scores title	reference	scored tempo	BEATS error
In the Wilderness	Lomax p. 91	112	correct
Lily Munroe	Lomax p. 164	126	correct
Yankee Doodle Dandy-O	Lomax p. 49	152	correct
Cumberland Gap	Lomax p. 162	232	correct
2/4 Default 5	new	103	default
Weeping Mary	Lomax p. 249	104	default
A Fair Beauty Bride	Lomax p. 195	108	default
2/4 Default 4	new	128	default
Whoa Back, Buck	Lomax p. 532	88	stretch
Come, Life, Shaker Life	Lomax p. 73	104	stretch
The Paw-Paw Patch	Lomax p. 92	104	stretch
Steal, Miss Liza	Lomax p. 501	138	stretch
Saturday Night	Lomax p. 499	80	+ phase
2/4 Positive Phase 4	new	95	+ phase
2/4 Positive Phase 3	new	107	+ phase
Brady	Sandburg p. 198	175	+ phase
Way Over in the Heavens	Lomax p. 248	50	- phase
The Blue-Tail Fly	Lomax p. 505	66	- phase
The Julie Plante	Lomax p. 127	76	- phase
The Kickin' Mule	Lomax p. 441	120	- phase

4/4 scores title	reference	scored tempo	BEATS error
The Greenland Whale Fishery	Lomax p. 61	84	correct
The Sow Took the Measles	Lomax p. 31	176	correct
I'm Bound to Follow the Longhorn Cows	Lomax p. 368	192	correct
Hudson River Steamboat	Lomax p. 85	200	correct
Hold the Wind	Lomax p. 474	88	default
Wade in the Water	Lomax p. 470	90	default
Careless Love	Lomax p. 585	100	default
4/4 Default	new	133	default
Shenandoah	Lomax p. 53	72	stretch
Frankie	Lomax p. 569	160	stretch
The Big Rock Candy Mountains	Lomax p. 422	168	stretch
Blow the Candle Out	Lomax p. 312	184	stretch
Hurray, Lie!	Lomax p. 260	120	+ phase
Shady Grove	Lomax p. 234	120	+ phase
Casey Jones	Lomax p. 564	200	+ phase
Single Girl	Lomax p. 166	200	+ phase
Dese Bones Gwine Rise Again	Lomax p. 476	100	- phase
The Gambling Suitor	Lomax p. 210	112	- phase
Sixteen Tons	Lomax p. 294	120	- phase
Buffalo Boy	Lomax p. 315	208	- phase

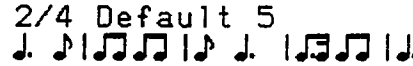
3/4 scores title	reference	scored tempo	BEATS error
The Longest Train	Lomax p. 541	104	correct
Run Along, You Little Dogies	Lomax p. 373	108	correct
The Horse Trader's Song	Lomax p. 323	156	correct
The Girl I Left Behind	Lomax p. 318	176	correct
I'm A-Leavin' Cheyenne	Lomax p. 378	126	default
Roll On, Columbia	Lomax p. 443	138	default
Sweet Betsy	Lomax p. 335	176	default
Let's Go A-Huntin'	Lomax p. 311	192	default
3/4 Stretch 10	new	98	stretch
3/4 Stretch 9	new	134	stretch
3/4 Stretch 8	new	139	stretch
3/4 Stretch 2	new	158	stretch
Irene	Lomax p. 593	152	+ phase
3/4 Positive Phase 11	new	154	+ phase
3/4 Positive Phase 6	new	157	+ phase
The Pretty Fair Widow	Warner p. 284	225	+ phase
3/4 Negative Phase 16	new	96	- phase
3/4 Negative Phase 15	new	118	- phase
3/4 Negative Phase 12	new	124	- phase
3/4 Negative Phase 13	new	135	- phase

6/8 scores title	reference	scored tempo	BEATS error
Lady Isabel and the Elf Knight	Lomax p. 18	120	correct
The Regular Army-O	Lomax p. 340	180	correct
I'ze the Bye	Lomax p. 149	228	correct
Lincoln and Liberty	Lomax p. 97	252	correct
Whoopee Ti Yi Yo	Sandburg p. 268	85	default
Plains of Baltimore	Warner p. 53	90	default
Negro Reel	Sandburg p. 134	140	default
The Sergeant, He is the Worst of All	Sandburg p. 435	140	default
6/8 Stretch 4	new	79	stretch
6/8 Stretch 5	new	85	stretch
6/8 Stretch 10	new	117	stretch
6/8 Stretch 9	new	137	stretch
6/8 Positive Phase 5	new	75	+ phase
6/8 Positive Phase 2	new	89	+ phase
6/8 Positive Phase 14	new	136	+ phase
6/8 Positive Phase 11	new	137	+ phase
The Mary L. Mackay	Lomax p. 144	99	- phase
Canada-i-o	Lomax p. 114	120	- phase
The British-American Fight	Warner p. 62	125	- phase
Root, Hog, or Die	Lomax p. 333	264	- phase

Scores of New Rhythms

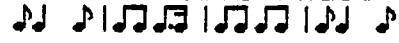
2/4 Default 4

2/4 Default 5

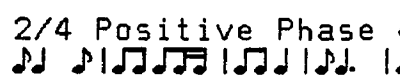


2/4 Positive Phase 3

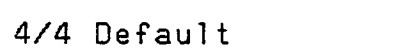
2/4 Positive Phase 4



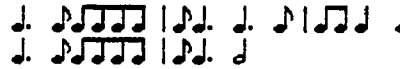
4/4 Default



3/4 Stretch 2



3/4 Stretch 8



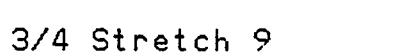
3/4 Stretch 9

3/4 Stretch 10

3/4 Positive Phase 6

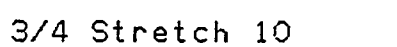
3/4 Positive Phase 11

3/4 Negative Phase 12



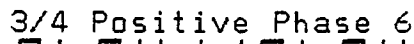
3/4 Negative Phase 13

3/4 Negative Phase 15

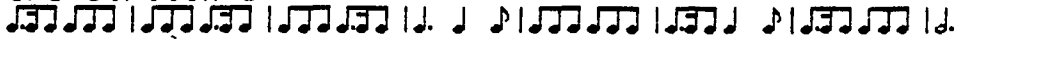


3/4 Negative Phase 16

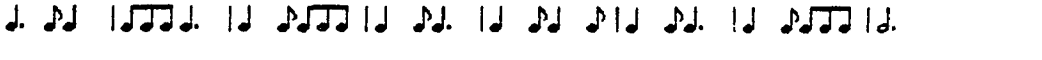
6/8 Stretch 4



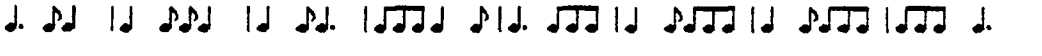
6/8 Stretch 5



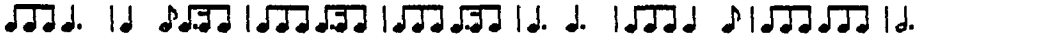
6/8 Stretch 9



6/8 Stretch 10

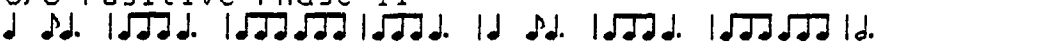


6/8 Positive Phase 2

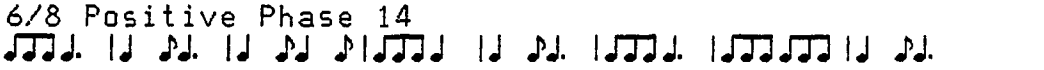


6/8 Positive Phase 5

6/8 Positive Phase 11



6/8 Positive Phase 14



References

Beauvillain, C. (1983). Auditory perception of dissonant polyrhythms. Perception & Psychophysics, 34, 585-592.

Bharucha, J. J. & Pryor, H. H. (1986). Disrupting the isochrony underlying rhythm: An asymmetry in discrimination. Perception & Psychophysics, 40, 137-141.

Brysbaert, M., Bovens, N., d'Ydewalle, G. & VanCalster, J. (1989). Turbo Pascal timing routines for the IBM microcomputer family. Behavior Research Methods, Instruments & Computers, 21, 73-83.

Collard, R., Vos, P. and Leeuwenberg, E. (1981). What melody tells about metre in music. Zeitschrift fur Psychologie, 189, 25-33.

Deliege, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's grouping preference rules. Music Perception, 4, 325-360.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. Perception & Psychophysics, 28, 381-389.

Deutsch, D. (1983). The generation of two isochronous sequences in parallel. Perception & Psychophysics, 34, 331-337.

Deutsch, D. & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. Psychological Review, 88, 503-522.

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

Fraisse, P. (1947-48). Mouvements rythmiques et arythmiques. L'Annee Psychologique, 47-48, 11-21.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), The Psychology of Music (pp. 149-180). New York: Academic Press.

Fodor, J.A. & Bever, T.G. (1965). The psychological reality of linguistic segments. Journal of Verbal Learning & Verbal Behavior, 4, 414-420.

Gregory, A.H. (1978). Perception of clicks in music. Perception & Psychophysics, 24, 171-174.

Handel, S. & Lawson, G.R. (1983). The contextual nature of rhythmic interpretation. Perception & Psychophysics, 34, 103-120.

Handel, S. & Oshinsky, J.S. (1981). The meter of syncopated auditory polyrhythms. Perception & Psychophysics, 30, 1-9.

Handel, S. (1984). Using polyrhythms to study rhythm. Music Perception, 1, 465-484.

Halpern, A.R. & Darwin, C.J. (1982). Duration discrimination in a series of rhythmic events. Perception & Psychophysics, 31, 86-89.

Jones, J.A., Miller, B.O. & Scarborough, D.L. (1988). A rule-based expert system for music perception. Behavior Research Methods, Instruments, & Computers, 20, 255-262.

Jones, J.A., Miller, B.O. & Scarborough, D.L. (1990). Discovering grouping structure in music. Proceedings of the Twelfth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jones, M.R. (1978). Auditory patterns: Studies in the perception of structure. In E.C. Carterette & M.P. Friedman (Eds.) Handbook of Perception (Vol. 8). New York: Academic Press, 255-288.

Jones, M.R., Maser, D.J. & Kidd, G.R. (1978). Rate and structure in memory for auditory patterns. Memory and Cognition, 6, 246-258.

Lashley, K.S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), Cerebral Mechanisms in Behavior. New York: Wiley.

Lerdahl, F. & Jackendoff, R. (1983). A Generative Theory of Tonal Music. Cambridge: MIT Press, 1983.

Lomax, A. (1960). The Folk Songs of North America. New York: Doubleday.

Longuet-Higgins, H.C. & Lee, C.S. (1982). The perception of musical rhythms. Perception, 11, 115-128.

Longuet-Higgins, H.C. & Steedman, M.J. (1971). On interpreting Bach. In D. Michie and B. Meltzer (Eds.), Machine Intelligence 6. Edinburgh: Edinburgh University Press.

Loy, G. (1985). Musicians make a standard: The MIDI phenomenon. Computer Music Journal, 9, 8-26.

Lunney, H.W.M. (1974). Time as heard in speech and music. Nature, 249, p. 592.

Marascuilo, L. A. (1971). Statistical Methods for Behavioral Science Research. New York: McGraw Hill.

Martin, J.G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. Psychological Review, 79, 487-509.

McNicol, D. (1972). A Primer of Signal Detection Theory. London: Allen & Unwin.

MIDI Manufacturers Association (1985). MIDI 1.0 Detailed Specification. North Hollywood, CA: International MIDI Association.

Mont-Reynaud, B. & Goldstein, M. (1985). On finding rhythmic patterns in musical lines. In Proceedings of the International Computer Music Conference, 1985. San Francisco: Computer Music Association, 391-397.

Oshinsky, J.S. & Handel, S. (1978). Syncopated auditory polyrhythms: Discontinuous reversals in meter interpretation. Journal of the Acoustical Society of America, 63, 936-939.

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

Palmer, C. & Krumhansl, C.L. (1987b). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. Perception & Psychophysics, 41, 505-518.

Perkins, D.N. (1974). Coding position in a sequence by rhythmic grouping. Memory & Cognition, 2, 219-223.

Povel, D.-J. (1981). Internal representation of simple temporal patterns. Journal of Experimental Psychology: Human Perception & Performance, 7, 3-18.

Povel, D.-J. (1984). A theoretical framework for rhythm perception. Psychological Research, 45, 315-337.

Povel, D.-J. & Essens, P. (1985). Perception of temporal patterns. Music Perception, 2, 411-440.

Povel, D.-J. & Okkerman, H. (1981). Accents in equitone sequences. Perception & Psychophysics, 30, 565-572.

Reber, A.S. (1973). Locating clicks in sentences: Left, center and right. Perception & Psychophysics, 13, 133-138.

Reber, A.S. & Anderson, J.R. (1970). The perception of clicks in linguistic and nonlinguistic messages. Perception & Psychophysics, 8, 81-89.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. Psychological Review, 77, 481-495.

Restle, F. (1972). Serial patterns: The role of phrasing. Journal of Experimental Psychology, 92, 385-390.

Rosenthal, D. (1989). A model of the process of listening to simple rhythms. Music Perception, 6, 315-328.

Sandburg, C. (1927). The American Songbag. New York: Harcourt, Brace & Company.

Scarborough, D.L., Jones, J.A., & Miller, B.O. (1989). Modelling music cognition: An expert system. The Arts & Technology II: Proceedings. Connecticut College, February 2-5, 1989.

Schulze, H.H. (1978). The detectability of local and global displacements in regular rhythmic patterns. Psychological Research, 40, 173-181.

Schulze, H.H. (1989). The perception of temporal deviations in isochronic patterns. Perception & Psychophysics, 45, 291-296.

Simon, H.A. (1968). Perception du pattern musical par AUDITEUR. Sciences de l'Art, V-2, 28-34.

Simon, H.A. & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. Psychological Review, 70, 534-546.

Simon, H.A. & Sumner, R.K. (1968). Pattern in music. In B. Kleinmuntz (Ed.), Formal Representation of Human Judgment. New York: Wiley.

Steedman, M.J. (1977). The perception of musical rhythm and metre. Perception, 6, 555-569.

Sternberg, S., Knoll, R. L. & Zukofsky, P. (1982). Timing by skilled musicians. In D. Deutsch (Ed.), The Psychology of Music. New York: Academic Press.

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

Sturges, P.T. & Martin, J.G. (1974). Rhythmic structure in auditory temporal pattern perception and immediate memory. Journal of Experimental Psychology, 102, 377-383.

Tenney, J. & Polansky, L. (1978). Hierarchical temporal gestalt perception in music: A "metric space" model. Privately circulated monograph.

Tenney, J. & Polansky, L. (1980). Temporal gestalt perception in music. Journal of Music Theory, 24, 205-241.

Turner, F. & Poppel, E. (1983). The neural lyre: Poetic meter, the brain, and time. Poetry, 142, 277-310.

Vorberg, D. & Hambuch, R. (1978). On the temporal control of rhythmic performance. In J. Requin (Ed.), Attention & Performance VII. NJ: Erlbaum.

Vos, P.G. (1977). Temporal duration factors in the perception of auditory rhythmic patterns. Scientific Aesthetics / Sciences de l'Art, 1, 183-199.

Warner, A. (1984). Traditional American Folk Songs from the Anne & Frank Warner Collection. Syracuse University Press.

Winer, B.J. (1971). Statistical Principles in Experimental Design. 2nd ed. New York: McGraw Hill.

White, B.W. (1960). Recognition of distorted melodies. American Journal of Psychology, 73, 100-107.