Journal of Experimental Psychology:
Human Perception and Performance
1980, Vol. 6, No. 3, 501-515

Recognition of Transposed Melodies: A Key-Distance Effect in Developmental Perspective

James C. Bartlett and W. Jay Dowling
University of Texas at Dallas

Four experiments examined the possibility of a key-distance effect in a transposition detection task. Subjects heard standard melodies followed by comparison melodies presented in the same key, a musically near key or a musically far key. The task was to recognize comparisons that were exact transpositions of the standards, rejecting nontranspositions. Results suggested a largely invariant key-distance effect with nontransposition comparisons (lures); same- and near-key lures evoked more false alarms than far-key lures. The variables of musical experience, age of subject, and familiarity of melody affected the level of transposition-recognition performance but did not consistently affect the size of the key-distance effect. The results support the psychological reality of key distance and are consistent with both musical and nonmusical-auditory theories of its effects. The key-distance effect was not found with transposition comparisons (targets), a result with implications for the separability of key and interval information in short-term memory for melodies.

Anyone with a good ear for music can sing, whistle, or hum familiar tunes correctly, that is, with the appropriate intervals among the notes. Yet, it is surprisingly difficult for nonmusicians to extract precise interval information from unfamiliar melodies on a single hearing. Attneave and Olson (1971) asked their subjects to transpose unfamiliar six-note melodies to new keys, using sine-wave oscillators. Their musically experienced subjects were able to do this, but untrained subjects were not. Cuddy and Cohen (1976) gave subjects the task of discriminating between an exact transposition of a novel three-note melody and a comparison melody in which one note had been changed. Musically trained subjects attained scores of nearly 90% correct on this task, but untrained subjects performed at about 60% correct. (Chance was 50%.) Dowling (1978) also gave subjects a transposition detection task and found chance discrimination between exact transpositions of novel melodies to new keys and tonal answers (i.e., comparison stimuli with the same contour—patterns of ups and downs—and in the same key as the standard stimuli but shifted in pitch and with different intervals among the notes). In all of these tasks with unfamiliar melodies, subjects seemed to have little trouble reproducing or recognizing the melodic contour, but they had a great deal of trouble with the exact-pitch intervals among the notes.

In exploring why subjects have difficulty with pitch intervals, we find it useful to distinguish among the various cues that might be used in judging the similarities and differences among pairs of melodies. Assuming that two tone sequences have the same number of notes, go at the same rate, and have the same rhythmic pattern, how can the listener tell them apart? He might use the set of pitches they contain, their contour, whether they are tonal or atonal, their keys if they are tonal, or whether they have the same intervals between notes.

This research was supported in part by Organized Research Funds granted by the University of Texas System.

The authors thank Jane Aten and the Dallas Independent School District for assistance in carrying out Experiment 4, and Darlene Smith for comments on an earlier draft of the manuscript.

Requests for reprints should be sent to James Bartlett, Program in Psychology and Human Development, University of Texas at Dallas, Box 688, Richardson, Texas 75080.

These types of cue vary considerably in how difficult they are to use and do not always operate independently of one another. It is relatively easy to discriminate between a repetition of a brief melody containing the same pitches as the standard stimulus and a distorted version having the same contour but different pitches (Dowling & Fujitani, 1971). It is also easy to distinguish between comparison melodies having the same contour as the standard and comparisons with different contour (Dowling, 1978; Dowling & Fujitani, 1971; White, 1960). It is somewhat harder to distinguish between tonal and atonal distortions of a tonal standard though not as hard for trained as for untrained subjects (Dowling, 1978). It is easier to detect distortions of intervals in tonal than in atonal melodies (Francès, 1958), but as noted earlier it is difficult to distinguish between the case in which the key has been changed but the intervals preserved (exact transposition) and the case in which the intervals have been changed but the key preserved (tonal answer). One theoretical account of this last result is that immediately after hearing an unfamiliar melody, the listener has available two classes of information: contour and key (the set of pitches of a modal scale anchored to a particular tonic).

According to this view there is little representation of specific interval sizes during early stages of learning, but in later stages representation of interval information improves, and listeners are able to recognize transpositions or to create them themselves. (This account applies to musically untrained listeners. Professional musicians doubtlessly extract interval information much more quickly.) There are two consequences of this account, which are explored in the present series of experiments. One is that with overlearned, familiar melodies, the scale steps of notes and the intervals between them should be well represented in memory, and the confusion between transpositions and nontranspositions should be reduced. This finding is already implied by the Attneave and Olson (1971) experiment; however, these authors sampled only one familiar melody (the National Broadcasting Company [NBC] chimes) and examined only transposition production, not the confusability between transpositions and different types of nontransposition in a recognition task. A second consequence is that key (e.g., C major, G major) might serve as a cue for melody equivalence independently of the intervals themselves. That is, two same-contour melodies should sound more similar, and more like transpositions, if they are played in the same or similar keys. This should hold regardless of whether the two melodies are actually transpositions or merely same-contour items with different interval patterns. This key-distance effect on recognition of transpositions and rejection of nontranspositions is investigated in all four experiments.

Key distance is well defined in western music, and that definition is used here. Two keys are considered closely related to the degree to which their diatonic modal scales share pitches. Thus, the keys of C major and G major are very close, since their scales (CDEFGABC and GABCDEF#G) share six out of seven different pitches. The scales of C-major and B-major (BC#D#EF#-G#A#B) are distantly related, since they share only two pitches (B and E). These relationships are coded in a convenient scheme called the "circle of fifths" in which two keys whose tonic centers are a fifth apart (i.e., for which the fifth note in one scale is the first note of the other) share all but one note. Two jumps around the circle of fifths gives keys that share all but two notes, and so forth. The circle of fifths orders all 12 major keys so that distance in either direction around the circle gives key dissimilarity in terms of shared notes. The sequence of keys is C-G-D-A-E-B-F# (or G^b)-D^b-A^b-E^b-B^b-F-C.

Although the major goal of these experiments is to examine the key-distance effect, we are also concerned with the effects of the age and experience of listeners. Musical training has typically been associated with good performance on transposition production and recognition tasks (Attneave & Olson, 1971; Cuddy & Cohen, 1976; Dowling, 1978), showing that sophisticated subjects are better at extracting information about muscial intervals. However, little is known concerning the effects of training on key distance and similarity. There are two conceivable outcomes. One possibility (suggested by Dowling, 1978) is that the more experienced subjects have more firmly internalized the tonal system of their culture and will therefore show a correspondingly greater effect of key distance. An alternate possibility is that since the system of keys in western music is based on similarity in terms of shared pitches, and the number of shared pitches is a purely physical variable that should need no cultural training to be effective, there should be little effect of experience on discrimination between keys. The finding of Dowling and Fujitani (1971) that even with atonal melodies the sharing of pitches was a powerful variable suggests that shared pitches are indeed important. Experiments 1, 2, and 3 manipulate experience as a dichotomous variable (more or less than 2 yr of training), and Experiment 4 is a replication of Experiment 3 with children of ages 5–10. This latter age range was chosen because work by Zenatti (1969) and others suggests that sensitivity to mode differences develops during this period.

One additional variable, presentation rate, is explored in Experiment 1. In one condition the 6 note/sec rate of Dowling (1978) is replicated, and in the other a much slower rate of 1 note/sec is used. We did not expect rate to interact with the key-distance effect but thought that it might interact with experience. That is, it seemed plausible that less experienced subjects might do relatively better at the slower rate at which their less practiced encoding skills would be effective.

General Method

Since these four experiments share many features, these will be described in a general method section. Table 1 shows the salient features of these experiments in comparison with those of Dowling and Fujitani (1971) and Dowling (1978).

Subjects

Undergraduates at the University of Texas at Dallas served in Experiments 1, 2, and 3 to fulfill a course requirement. The subjects' mean age was 29.2 yr, and their mean amount of musical training was 3.5 yr, defined as lessons on an instrument or voice or participation in an instrumental ensemble. For purposes of data analysis, subjects were dichotomized into inexperienced (less than 2 yr of training) and experienced (2 yr or more of training) groups. Subjects with some training but less than 2 yr were rare, and the mean amount of training of the inexperienced and experienced groups was .17 yr and 6.8 yr, respectively. Many of the subjects had served in similar music recognition experiments during the preceding year. The subjects served in group sessions, and the same group of subjects served in both Experiments 2 and 3. Approximately 65% of each group in each experiment was female.

Stimuli

The stimuli were produced on a freshly tuned Steinway piano, tape recorded, and presented to subjects over loudspeakers at comfortable listening levels. Timing was controlled by a Davis timer, which emitted barely audible clicks at intervals of .33 sec or 1.00 sec (Experiment 1) or .67 sec (Experiments 2, 3, and 4). The interstimulus interval (ISI) was kept constant throughout all four experiments and was 4.0 sec. This ISI was longer than the 2.0-sec interval used by Dowling (1978) so that the pair of stimuli in the slow condition of Experiment 1 might be separated sufficiently so as not to run together.

Procedure

Subjects were instructed that this was an experiment on memory for melodies. On each trial of the experiment, they heard a pair of melodies. The subjects' task was to judge whether the two melodies were the same or different and to respond using a four-level confidence scale. Subjects wrote their responses sequentially on a sheet of paper. Separate sheets were used for the two conditions of Experiment 1. Subjects had 6.0 sec to respond on each trial, and the onset of each trial was announced by the experimenter's voice saying the trial number 2.0 sec before the start of the standard stimulus.

Before the start of each session, the subjects listened to examples of each trial type used in the experiment, constructed out of familiar melodies not used in the rest of the experiment. The reasons each type of lure was different from the standard were explained, and the experimenter emphasized that the subjects were to respond *same* only to exact transpositions. The set of examples began and ended with an example of exact transposition.

Experiment 1

In Experiment 1 standard stimuli were followed by six types of comparison stimuli (see Table 1): transpositions (T); tonal lures in the same key as the standard (LS),

Table 1
Comparisons of Stimulus Conditions

	Standard stimulus					Comparison stimulus						
Experiment	Generation	Start pitch or key	Interval size (semitones)	Duration of common note (sec)	ISI (sec)	Types	Reference note[a]	Keys[b]	Scale step of ref. note[a]	Interval change[c]	Note change inferred[d]	Note change actual (%)[e]
Dowling & Fujitani (1971) (transposed conditions)	Atonal random walk	C	1.75	.17	2.0	T	±1–7 semitones					
						At						
						D						
Dowling (1978)	Tonal random walk	C	2.3	.17	2.0	T	A or E	A,E	1	0	1.8	
						LS		C	6,3	2.4	0	
						At		—	—	2.4	1.8	
						D		A,E	1	—	1.6	
Experiment 1	Tonal random walk	C	2.3	.17 or 1.0	4.0	T	A or E	A,E	1	0	2.3	66
						LS		C	6,3	2.9	0	52
						LN		G,D	2	2.6	1.1	64
						LF		E,B	4	2.3	3.0	89
						At		—	—	2.2	2.0	74
						D		A,E	1	—	2.3	78
Experiment 2	Familiar melodies & tonal constructs	C	2.3	.67	4.0	TN	±1–6 semitones	var.	1	0	0	88
						TF		var.	1	0	3	83
						LS		C	var.	3	0	47
						LF		var.	var.	3	3	80
Experiments 3 & 4	Familiar melodies	all 12	2.3	.67	4.0	TN	±1–2 semitones	B^{b},D	1	0	1.8	45
						TF		B,D^{b}	1	0	4.0	91
						LS		C	2,7	3.4	0	32
						LN		A,D	2,7	3.6	1.8	41
						LF		B	2,7	3.8	3.6	76

Note. T means exact transpositions of the standard stimuli; TN means transpositions to near keys; TF means transpositions to far keys; L refers to tonal stimuli that share contour but not intervals with standard stimuli in the same (LS), near (LN), or far (LF) keys relative to the key of the standards; At means atonal stimuli that share contour with the standard stimuli; D means stimuli with different contours from standard stimuli.

[a] By reference note we mean the note that was the tonic of the standard stimulus (e.g., C if the standard was in the key of C major). All unfamiliar standards used in Experiments 1 and 2 began on the tonic; thus, the starting note was the same note as the reference note for those stimuli. However, in some of the familiar melodies, the tonic occupied a later temporal position and therefore the second, third, or fourth note might have been the reference note. In all standards the scale step of the reference note was, by definition, one. Likewise, in any transposition, the scale step of the reference note was one. However, in nontranspositions the scale step of the reference note was never one. For example, if a standard (in the key of C) began as C, G, E, an LS item might begin as A, E, C. The scale step of the reference note (the starting note in this case) is one in the case of the standard and six in the case of the comparison.

[b] Relative to C major. In Experiments 3 and 4, if the key of the standard was F, then B^{b} in this column denotes E^{b}, D denotes G, and so forth.

[c] Interval changes from the standard were counted as types (not tokens).

[d] Notes in comparison not in inferred set (scale) of standard were counted as types.

[e] Notes in comparison not actually in the standard were counted as tokens (not types) and expressed as the mean percent of total notes for stimuli within the given category.

in a nearly related key (LN), or in a far key (LF); atonal lures with the same contour as the standard (At); and different contour tonal lures (D). The Interval Change column in Table 1 gives the number of intervals in the comparison that were different from the corresponding intervals in the standard, counted as types, not tokens. That is, if an interval were repeated, it was only counted once. In the sense used here, the first phrase of "Mary Had a Little Lamb" contains two intervals, the first phrase of "Frère Jacques" contains three intervals, and the first phrase of "Twinkle Twinkle Little Star" contains two intervals. Ascending and descending intervals between the same pitches (e.g., do-re and re-do) were counted only once. Unisons were disregarded. The Note Change Inferred column in Table 1 gives the number of pitches in the comparison not contained in the diatonic scale of the standard, for example, an F# when the standard was in the key of C. These were counted as types, so that repeated notes were counted only once. The Note Change Actual column in Table 1 gives the percentage of notes in the comparison that did not appear in the standard and were counted as tokens. In Experiment 1, only the zero entries under Interval Change for T stimuli and under Note Change Inferred for LS stimuli were controlled directly. The other values varied as a consequence of the operation of the other constraints imposed on the stimuli. Note, for example, that among tonal lures the farther removed the key, the more actual and inferred note changes there are.

Method

The two conditions of Experiment 1 each consisted of 72 trials. In the fast condition, the duration of a quarter note (see Figure 1) was .17 sec and in the slow condition 1.0 sec. Each standard and comparison stimulus consisted of five notes of which the fifth had twice the duration of the first four. Thus, stimuli were 1.0 sec long in the fast condition and 6.0 sec long in the slow.

Stimuli were constructed in the same way as described by Dowling (1978). Each of the 16 possible contour patterns of five-note melodies appeared in a standard stimulus four or five times in the 72 trials of a condition. Contours were randomly assigned to trial types, and trial types were randomly assigned to trial numbers within four blocks of 18 trials. Within each block all trial types were as equally represented as possible. All standard stimuli started on middle C ($F_0 = 262$ Hz, where F stands for fundamental frequency) and were in the key of C major. The melodies were generated as a random walk on the tonal scale, using the randomly selected contour and probabilities of diatonic scale step sizes of $P(\pm 1 \text{ step}) = .67$ and $P(\pm 2 \text{ steps}) = .33$. This gave an expected interval size of 2.3 semitones, which is roughly the same as the mean interval size for the familiar tunes used in the other experiments.

Figure 1 shows examples of the different types of comparison stimuli used in Experiment 1. In the 72 trials, there were 18 transpositions, 12 tonal lures in the same key, 9 in a near key, 9 in a far key, 12 atonal, and 12 different contour. Half of the comparisons began on A below middle C, the other half on the E above, randomly determined and balanced within

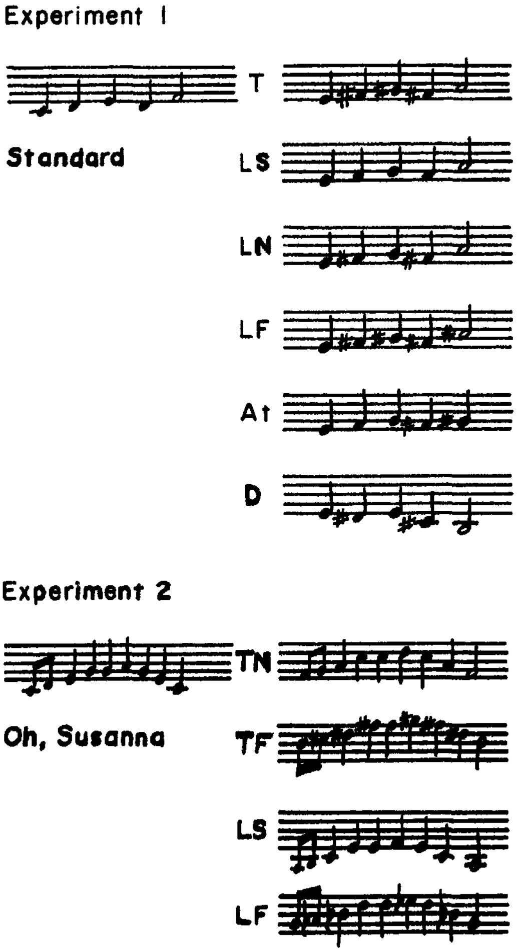

Figure 1. Examples of stimuli used in Experiment 1 (top) and Experiment 2 (bottom).

each trial type. T stimuli were simply transposed to the key of A or E and so began on the first scale step or tonic of those keys. (See the column labeled Scale Step of Reference in Table 1). T stimuli retained both the diatonic and physical interval sizes of the standards. LS stimuli stayed in the key of C major and so changed physical interval sizes measured in semitones from the standards but not diatonic interval sizes. They began on either the sixth (A) or third (E) scale step of C major. For LN stimuli, A and E became the second scale steps of the keys of G and D, both nearly related to C. For LF stimuli, A and E were the fourth scale steps of the keys of E and B, both relatively distant from C.

The At stimuli were generated using the same contours as the standards but with an interval distribution that introduced nondiatonic intervals while retaining the same expected interval size of 2.3 semitones: $P(\pm 1$ semitone$) = .17$; $P(\pm 2$ semitones$) = .33$; $P(\pm 3$ semitones$) = .50$. The constraint was added (and not present in Dowling, 1978), that not only should intervals be randomly selected but the resulting tone sequences should be impossible in any western mode, including the harmonic and melodic minors.

Different contour stimuli were generated in the same way as standards, with two out of the four interval directions randomly selected for change from the contour of the standard and with the constraint that no contour appear more than once as a D comparison stimulus in the 72 trials. D stimuli were in the keys of A or E and thus began on the first scale step in those keys. Changes in interval size are irrelevant and not tabulated because of the changes in direction.

The fast and slow conditions used the same stimulus pairs on like trial types, but the trials were presented in different random orders in the two conditions. Twenty-four subjects served in Experiment 1, and all subjects did both the fast and the slow conditions in separate sessions separated by a 5-min break. Fourteen subjects did the conditions in the order fast–slow, and 10 subjects, in the reverse order. Half of the subjects with each order were experienced and the other half inexperienced.

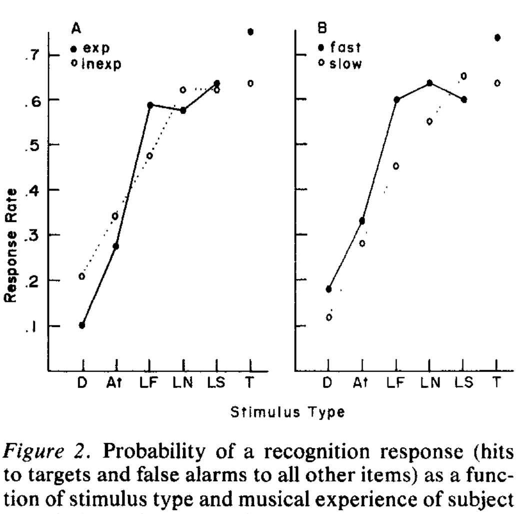

Figure 2. Probability of a recognition response (hits to targets and false alarms to all other items) as a function of stimulus type and musical experience of subject (panel A) and presentation rate (panel B). (exp = experienced; inexp = inexperienced.)

Results

For scoring purposes the four confidence levels were collapsed into the two categories, same and different, with two response types going into each category. The false alarm data (i.e., responses of *same* to any but the T stimuli) were fed into a four-way analysis of variance with lure type (LS, LN, LF, At, D), experience, presentation rate (fast, slow), and test order (fast-slow, slow-fast) as factors. The largest effect was that of lure type, $F(4, 80) = 67.24$, $p < .001$, $MS_e = .029$. The other significant main effect was that of rate, $F(1, 20) = 6.21$, $p < .05$, $MS_e = .024$, with fewer false alarms at the slow rate. There were significant interactions between Lure Type × Rate, $F(4, 80) = 2.75$, $p < .05$, $MS_e = .022$; lure type × experience, $F(4, 80) = 2.95$, $p < .05$, $MS_e = .029$; and Experience × Rate × Order, $F(1, 20) = 5.74$, $p < .05$, $MS_e = .024$. No other effects approached significance except for a marginal interaction between Experience × Rate, $F(1, 20) = 4.11$, $p < .06$, $MS_e = .024$.

Figure 2 displays the most important among these results. In both panels of Figure 2, the main effect of lure type can be seen clearly, with D and At stimuli producing fewer false alarms than the tonal lures. Planned comparisons between adjacent pairs of lure types showed significant differences at the .05 level for D versus At, $t(23) = 4.15$, At versus LF, $t(23) = 5.61$, LF versus LN, $t(23) = 2.11$, but not for LN versus LS, $t(23) = .77$. A post hoc comparison showed a highly significant difference for LF versus LS, $t(23) = 3.41$, $p < .001$. To determine the generalizability of these effects over items, we did a parallel set of t tests over items rather than over subjects. All of the foregoing significant effects were supported at the .05 level except for LF versus LN. Thus, the results support a key-distance effect; nontranspositions in far keys (LF items) were easier to reject than tonal answers (LS items) and

Table 2
Probabilities of Responding Same *to Targets and Lures in Experiment 1*

		Targets		Lures	
Experience	Order	S	F	S	F
Inexperienced	F–S	.66	.70	.42	.50
	S–F	.55	.59	.48	.42
Experienced	F–S	.71	.80	.41	.47
	S–F	.63	.82	.37	.49

Note. S = slow; F = fast.

possibly easier to reject than nontranspositions in near keys (LN items).

The Lure Type × Experience interaction shown in Figure 2A was analyzed by five *t* tests, comparing experienced to inexperienced subjects on each type of lure. Of these five tests, only two indicated a reliable experience effect, D, $t(22) = 2.41$, and LF, $t(22) = 2.33$. Note that these differences are in opposite directions—Experienced subjects made more false alarms to LF items and fewer false alarms to D items. Experienced subjects also made significantly more hits to transpositions than inexperienced subjects, $t(22) = 2.20$.

Figure 2B displays the effect of presentation rate and the Lure Type × Rate interaction. The generally lower false alarm rates at the slow rate carry over to lower hit rates to T stimuli. This suggests that the main effect of rate is due to a criterion shift and not to greater accuracy. This supposition is borne out by the lack of any significant rate effect in the analysis of area scores reported later. We explored the Lure Type × Rate interaction with five *t* tests, one for fast versus slow rates for each type of lure. Only the test for LF items showed a reliable effect, $t(23) = 3.26$. Thus, the data suggest that the key distance effect is stronger for slow than for fast stimuli and stronger with inexperienced than experienced subjects.

The three-way interaction of Experience × Rate × Order was unexpected and is puzzling. This interaction is represented in Table 2 along with corresponding data on hits. One reasonable description of the pattern is as follows: Fast presentation rate was always associated with both higher false alarm rates *and* higher hit rates, except with inexperienced subjects in the slow–fast test order. We attribute this pattern to a practice effect. For inexperienced subjects practice with slow melodies facilitates accurate performance on the later test with fast melodies, whereas practice with fast melodies does not help much.

Area scores. In the present experiments, we have found response rates reported in terms of hits and false alarms such as those analyzed earlier to be more informative than discrimination measures such as area under the memory-operating characteristic (MOC; Swets, 1973). However, since Dowling (1978) reported his data in terms of area scores, we also show the present data in that form for comparison. Table 3 shows area under the MOC calculated using the four categories of confidence level responses, comparing hit rates to T items with each of the five false alarm rates. Comparison of the means in rows 3 and 6 of Table 3 shows that the present results conform to those of Dowling. Dis-

Table 3
Areas Under the Memory-Operating Characteristic for Experiment 1 Compared With Those of Dowling (1978)

Study and experience	T vs. LS	T vs. LN	T vs. LF	T vs. At	T vs. D
Dowling (1978)					
I	.49	—	—	.59	.81
E	.48	—	—	.79	.84
M	.49	—	—	.69	.83
Experiment 1					
I	.50	.48	.58	.66	.76
E	.57	.57	.62	.78	.87
M	.54	.53	.60	.72	.82

Note. I means inexperienced; E means experienced; T means exact transpositions of the standard stimuli; L refers to tonal stimuli that share contour but not intervals with standard stimuli in the same (LS), near (LN), or far (LF) keys relative to the keys of the standards; At means atonal stimuli that share contour with the standard stimuli; D means stimuli with different contours from standard stimuli.

crimination between T and LS items is near chance, whereas discrimination between T items and both At and D items is substantially above chance. These results also show that discrimination between T and LN items is about the same as that between T and LS items, whereas the T versus LF discrimination is somewhat better. The area scores for LN and LF items differed significantly by a *t* test, $t(23) = 4.56$, $p < .01$.

The effects of experience differed in the two experiments. Whereas Dowling (1978) found a sizable experience effect with only At items, *t* tests on the present area-under-MOC data indicated significant ($p < .05$) experience effects for every item type except LF. Experiments 2 and 3 provide further looks at experience effects. It should be noted that the pattern of experience effects in Table 3 (area scores) are consistent with those indicated by the "raw" response probabilities shown in Figure 2. In Figure 2 it can be seen that more experienced subjects showed higher hit rates than less experienced subjects but did not show higher false alarm rates, except with LF items. Thus, experience affected discrimination between targets and between each type of lure except LF items.

Discussion

The most important finding of Experiment 1 was that LF items—same-contour tonal-comparison items in a far key from the standard—are harder to reject than LS items and probably LN items as well. Since LF items are tonal, their relatively high rejection rate cannot be due to a strategy of rejecting atonal items. If these LF items sound different from their standard stimuli, it is because of the relatively distant relationship of their keys and those of the standards. If key distance is responsible for perceptual dissimilarity between LF lures and standards, then it should also affect the similarity of targets in near and far keys. Experiments 2, 3, and 4 introduce a distinction between TN and TF stimuli, that is, between targets transposed to near and far keys, respectively. The key-distance effect should lead hit rates to be lower for TF items than for TN items because of their dissimilarity to the standard stimuli.

Experiments 2, 3, and 4 test the replicability of the key-distance effect with familiar melodies. One expectation we had was that transpositions would be much more accurately recognized with familiar melodies, since the intervals of those melodies would be thoroughly overlearned. We expected this effect of familiarity to appear with inexperienced subjects as well as with experienced ones, as suggested by the success of Attneave and Olson's (1971) inexperienced subjects in reproducing the intervals of the familiar NBC chimes. An effect of familiarity on the size of the key-distance effect was considered an open question. On the one hand, it seemed possible that subjects could effectively separate interval information from key information when hearing a tune with a previously established representation in long-term memory. If so, key distance would have no effect with familiar melodies. On the other hand, familiarity might simply improve the extraction of interval information without affecting its separability from key information. In this latter case, key-distance effects should be found with both familiar and unfamiliar materials (unless precluded by ceiling or floor effects).

Experiment 2 was a test of the key-distance effect with both near- and far-key targets and lures. It was a simplification of Experiment 1 in that we did not think further replication of the ease of rejection of At and D items was necessary. Along with this simplification we decided to control directly the number of interval changes in lures and the number of inferred note changes in far-key items. This is shown in Table 1, which also indicates those cases in which key or scale step of reference was allowed to vary to contrive exactly three changes in interval or note. Examples of the types of stimuli used in Experiment 2 are shown in Figure 1.

Experiment 2 also contrasted the recognition of first phrases of familiar melodies with recognition of unfamiliar stimuli. The unfamiliar melodies were constructed so that each had the same rhythmic pattern as one of the familiar melodies but a different melodic pattern. Due to contraints on stimulus construction (see later), it was impossible to assign melodies randomly to trial types. Therefore, every effort was made prior to Experiment 2 to assign stimuli so that melodies in each trial type would be of roughly equal familiarity. As a final check, we had the subjects of Experiments 2 and 3 rate each of the familiar melodies for familiarity at the end of the experimental session. Mean percent familiarity judgments ranged from 88 to 96 over the four trial types. We searched thoroughly for correlations, both across items and across subjects, for effects of familiarity on recognition performance and found none.

Experiment 2

Method

There were 40 trials in Experiment 2. There were four types of trials (TN, TF, LS, and LF), and there were five instances of each trial type with both familiar and unfamiliar melodies. The order of the 40 trials was randomized with the constraint that familiar melodies and their same-rhythm unfamiliar counterparts not appear consecutively. A given familiar melody and its unfamiliar mate both occurred in the same trial-type category. The duration of the most frequent note in each of these melodies was set at .67 sec—a quarter-note duration in Figure 1. The modal stimulus was 5.33 sec long, with a range of 2.67 sec–9.33 sec. The familiar melodies were selected on the basis of subjective familiarity to subjects in prior research. All were in major keys.

The assignment of familiar melodies to trial types was constrained by logical possibilities available with each melody. For example, since the first phrase of "Twinkle Twinkle Little Star" contains just two intervals (counted as types), it could not appear as an LS or LF stimulus, since those categories require three interval changes (see Table 1). It should be noted that since TN stimuli could not contain any pitches not implied by the C-major scale of the standard, they could have been called TS if that category were not logically impossible in music theory. Construction of the yoked-control unfamiliar melodies was constrained by the same need to use the melody in the assigned trial category and also by the requirement of producing an overall mean interval size of 2.3 semitones. Apart from those constraints, the unfamiliar melodies were simply constructed to sound alright—that is, to be reasonably good gestalts. No attempt was made to randomize note or interval selection. A little experimentation with the constraints listed or implied earlier will convince the reader that such an attempt would have boggled the mind of a J. S. Bach or a large computer. But the result was a set of unfamiliar stimuli that was probably more musically meaningful than the stimuli of Experiment 1.

Fifteen subjects served in Experiment 2—7 inexperienced and 8 experienced.

Results

Table 4 gives the proportions of *same* responses to targets and lures in Experiment 2, collapsed across confidence levels as in Experiment 1. A four-way analysis of variance was performed on these data, with the factors of familiarity, interval change (target vs. lure), key change, and experience. The only significant main effect was that of interval change, $F(1, 13) = 273.0, p < .001$, $MS_e = .035$, indicating successful discrimination between targets and lures. There was a significant interaction between Interval Change × Experience, $F(1, 13) = 8.89, p < .02$, $MS_e = .035$, supporting the tendency for experienced subjects to outperform inexperienced subjects. There was also a significant interaction of Interval Change × Familiarity, $F(1, 13) = 46.4$, $p < .001$, $MS_e = .023$, supporting the predicted trend for better interval discrimination with familiar melodies. Of most interest, there was an Interval Change × Familiarity × Key Change interaction, $F(1, 13) = 4.98$, $p < .05$, $MS_e = .029$. This interaction is clear in rows 3 and 6 of Table 4, collapsing over experience. It can be attributed to the

Table 4
Probabilities of Responding Same *to Targets and Lures in Experiment 2*

Familiarity and experience	Targets		Lures	
	TN	TF	LS	LF
Unfamiliar				
I	.69	.74	.49	.29
E	.70	.75	.40	.20
M	.70	.75	.45	.25
Familiar				
I	.97	.86	.29	.34
E	1.00	1.00	.15	.03
M	.99	.93	.22	.19

Note. I means inexperienced; E means experienced; TN and TF refer to transpositions to near and far keys, respectively; LS and LF refer to same-contour imitations in same and far keys, respectively.

relatively high false alarm rates to LS items with unfamiliar melodies. In fact, of the four comparisons testing for the key-distance effect within the categories of familiar–unfamiliar and target–lure, only the test between unfamiliar LS and LF items was significant, $t(14) = 2.63$. None of the four t tests was significant when performed over items rather than subjects, but this can be attributed to the small number of items (five) used in each category. Since Experiments 1 and 3 both support the key-distance effect even with items taken as the random factor, we are confident that the key-distance effect generalizes over items.

Discussion

Experiment 2 replicated the key-distance effect, but only in terms of false alarm rates to unfamiliar items. Experiment 3 was designed as a further test of the key-distance effect concentrating on just familiar melodies. Experiment 2 also demonstrated clearly that interval changes in familiar melodies are much easier to detect than interval changes in unfamiliar ones. Both experienced and inexperienced subjects appear to have melodic interval sizes stored in their long-term memories for those tunes they know, and those interval sizes can be retrieved in such a way as to help them detect changes of interval size when they occur.

Experiment 2 replicated the positive effect of musical experience on performance; experienced subjects were better at the transposition detection task. However, the qualitative nature of this experience effect differed in the two experiments. In Experiment 1, the more experienced subjects tended to show a smaller key-distance effect, whereas in the present study, the key-distance effect was, if anything, larger for more experienced subjects. (Only experienced subjects showed a trend toward a key-distance effect with familiar lures.) Experiments 3 and 4 provide additional information on this point.

Experiment 3

Experiment 3 was conducted with the same subjects as Experiment 2—It simply followed in the same session after a 5-min break. Experiment 3 had two purposes. First, it provided a more extensive test of the key-distance effect with familiar melodies, using more melodies and adding LN items between LS and LF as in Experiment 1.

A second purpose of Experiment 3 was to provide a task that could be used to compare the performance of children and adults with respect to the key-distance effect. Pilot work showed that the unfamiliar materials of Experiment 1 were much too difficult for the children to disclose differences due to detection of interval changes. We therefore constructed a set of stimuli using familiar melodies, which we could use in parallel experiments with adults and children, namely, Experiments 3 and 4.

Experiment 3 incorporated some methodological refinements, which can best be described with reference to Table 1. In Experiment 3 the reference scale steps to which the tonal lures were moved were controlled precisely, and the shift in pitch of the starting notes of both targets and lures was carefully counterbalanced across key distance and item type. In Experiment 1, reference scale step was completely confounded with key distance. That may have been an innocuous confounding, but it is conceivable that shifts of comparisons to the fourth degree of the scale make interval changes easier to detect than shifts to the third or sixth degree of the scale. With the randomly generated materials of Experiment 1, this confounding was a virtual necessity—or at least the set of rules necessary to eliminate the confounding and still preserve random generation would have been very long. In Experiment 2 switching to familiar tunes and yoked-control constructions took us away from the complexities of random generation, and we chose to control the number of interval changes and the number of note changes precisely. The price we paid for that was that we had to use whatever key and reference scale step would work with the comparison stimulus. At the very least this introduced noise into the data and at worst might have involved unknown confounds that we would have no means of assessing.

In Experiment 3 we exerted precise control over the keys and reference scale steps of comparison stimuli and monitored the resulting pattern in terms of interval changes and note changes. Note changes checked out reasonably well in that there were substantially more changes for far keys than for near keys, and the values of TN and LN, and TF and LF, were close. For interval changes, which we wanted to hold constant for all lure stimuli, the range across item types was 3.4–3.8 changes, which was better than the range in Experiment 1 of 2.2–2.9 changes.

Method

There were 25 trials in Experiment 3: 5 each of 5 trial types. The trial types were TN, TF, LS, LN, and LF. All standard stimuli were familiar melodies having the same timing characteristics as in Experiment 2, except that here the first two phrases of each melody were used. The modal stimulus was 10.67 sec long, and the range was 5.33–16.0 sec. (Two phrases were used to make the task easier for the children of Experiment 4). At the end of the session, the subjects rated the familiarity of all 25 tunes used in the experiment. They were presented again with each tune and were asked to give its title or some of the words or name its appropriate occasion, or failing that to state whether it sounded familiar. The results of these two kinds of response were closely parallel, so only the percentage of subjects responding at the looser criterion will be reported. The mean percent of subjects familiar with the tunes used for the TN and TF categories were 72% and 79%, respectively, and for the LS, LN, and LF categories, 94%, 91%, and 89%, respectively. The a priori attempt to equalize familiarity across key distance was therefore successful, and the decision to bias familiarity, if it were to be biased at all, against targets was also successful.

Unlike Experiments 1 and 2, the standard stimuli in Experiment 3 were randomly assigned to all 12 different keys, with the constraint that all keys appear at least twice. For simplicity of exposition, we will describe the changes of key of the comparison stimuli relative to C major. (This description is as though we constructed each trial in C major to begin with and then randomly transposed the various trials to different keys.) The reference note or tonic of the comparison melody was always one or two semitones above or below C; that is, B^b (A#), B, D^b (C#), or D. (We say "reference note" rather than "starting note" because not all familiar tunes start on the first degree of the scale.) Altogether five stimuli were shifted to D (+2 semitones), eight to D^b (+1 semitone), seven to B (−1 semitone), and five to B^b (−2 semitones). Of those shifted to D, three were TN and two were LS. Of those shifted to Db, three were TF, three were LN, and two were LF. Of those shifted to B, two were TF, three were LS, and two were LN. Of those shifted to B^b, two were TN and three were LF. For transpositions the two near keys were B^b and D major, both of which share five-sevenths of their pitches with C major. The far keys were B and D^b major, both of which share two sevenths of their pitches with C major. Note that the TN stimuli, whose scale step reference is one, simply shifted to B^b or D; and the TF stimuli, again referenced to the first scale step, shifted to B or D^b. The LS stimuli stayed in C major and shifted to either the second (D) or seventh (B) scale step. The LN stimuli went into either A or D major, using the second (B) or seventh (C#) scale step, respectively. The LF stimuli went into B major, in which A# is the seventh scale step and C# the second.

Results

Table 5 shows proportions of positive responses as a function of item type and experience, collapsed into two response categories as in the analysis of Experiments 1 and 2. As in Experiment 2 discrimination between targets and lures with familiar materials was good, with uniformly high hit rates for both groups of subjects, and so we did an analysis of variance on just the false alarm data. (In passing, we note that there is no evidence for a key-distance effect with the targets, but this may be due to ceiling effect.) In this 3 × 2 (Lure Types × Levels of Experience) analysis, the main effect of lure type was significant, $F(2, 26) = 3.52$, $p < .05$, $MS_e = .017$, and represented essentially a key-distance effect. A t test over items supported the difference between LS and LF items, $t(8) = 2.69$,

Table 5
Probabilities of Responding Same *to Targets and Lures in Experiments 3 and 4*

	Targets		Lures		
Experience & age	TN	TF	LS	LN	LF
Inexperienced adults	.89	.89	.23	.11	.03
Experienced adults	.93	.95	.05	.05	.00
M	.91	.92	.14	.08	.02
Kindergarten	.67	.58	.65	.57	.55
Grades 1 & 2	.58	.54	.60	.60	.36
Grade 3	.70	.62	.53	.48	.35
M	.65	.58	.59	.55	.42

Note. LN refers to same-contour imitations in near keys. TN and TF refer to transpositions to near and far keys, respectively; LS and LF refer to same-contour imitations in same and far keys, respectively.

$p < .05$. This result contrasts with the nonsignificant key distance effect with familiar tunes of Experiment 2. The effect of experience approached significance, $F(1, 13) = 3.91$, $p < .07$, $MS_e = .023$. The interaction had an F ratio close to 1.

Experiment 4

Experiments 1, 2, and 3 have clearly demonstrated a key-distance effect in the detection of interval changes in transposed melodies, at least indicating higher false alarm rates when tonal lures are closer to the key into which the lure has been transformed. The key-distance effect seems to be a robust phenomenon not consistently dependent on the familiarity of the muscial stimuli or the muscial training of the listeners. Experiment 4 explores the possibility that the key-distance effect might be largely independent of the age of the listeners. Children of ages 5–8 performed the transposition detection task with the same stimuli used in Experiment 3. This age range seemed most appropriate to investigate, since previous studies had found important changes in performance on melody recognition tasks during that period. Zenatti (1969) had children say which note of a melody had been changed in pitch. She found that with three-note melodies the average success rate on this rather difficult task went from 25% (around chance) at age 5 to about 50% at age 9. Imberty (1969) has shown that already by the age of 8, children are sensitive to changes of mode (major vs. minor) and key when these are introduced in the midst of a somewhat familiar melody.

Method

Experiment 4 used the same materials as Experiment 3. The main differences in procedure were that subjects were run individually and the experimenter recorded their responses. Only the responses *same* and *different* were used, since pilot attempts to use the four-response system with kindergarteners ran into problems.

Three groups of subjects participated. There were 23 kindergarten students of mean age 5.6 yr at the time of testing, 11 first and second graders of mean age 6.9 yr, and 12 third graders of mean age 8.6 yr. All subjects had participated for the better part of 1 school yr in federally funded music enrichment programs for inner-city children and so were if anything more sophisticated musically than their peers. (This sophistication is substantiated by tests of melodic memory with similar groups.) Although the groups were by definition selected for academic achievement below national norms, the research just cited shows substantial gains in reading scores by children in these enrichment programs, and so we believe a fair characterization of these subjects is that they were probably about average in academic performance for their ages.

Results

Table 5 shows proportions of *same* responses to the stimuli of Experiment 4 by children of different ages. An analysis of variance of the hit data showed no effects. For a description of overall discrimination, rather than area under the MOC (which for two-category data would be rather unstable), we did an analysis of A' (Grier, 1971) data. This showed only an effect of age, with the older subjects performing better, $F(2, 43) = 4.18, p < .05, MS_e = .007$. (In calculating A', TN was compared with LN, and TF with LF, LS being dropped from the analysis.) The analysis of variance on false alarm data showed a significant key-distance effect, $F(2, 86) = 9.18$, $p < .001$, $MS_e = .037$. There were no other significant effects. (A t test over items for LS vs. LF gave a significant difference by a one-tailed test, $t(8) = 1.91$.) Notice that with the youngest children there was only a key-distance effect, with no evidence of discrimination. This is consonant with the findings of Riley, McKee, Bell, and Schwartz (1967) that for children of this age, the pitch of tones in a pair is much more salient than the interval between them.

General Discussion

Figure 3 shows false alarm data from all four experiments. The main point of interest is that the key-distance effect is apparent in all sets of data. The key-distance effect did not appear as we expected in hit data, but it appeared in false alarm data across age and experience without qualitative differences. In fact the performance of adults with unfamiliar melodies (Experiment 1) closely parallels the performance of children with familiar melodies (Experiment 4). This constant effect of key distance on false alarms is the most important finding of these studies. Although the key-distance effect did not occur for some subjects in certain conditions (for experienced subjects with unfamiliar materials in Experiment 1 and for inexperienced subjects with familiar tunes in Experiment 2), the accumulated data suggest that the key-distance effect is not consistently dependent on age or musical training. Training, whether of the formal sort that differentiates our experienced subjects from the inexperienced or of the informal sort that leads people to know a few familiar tunes of their culture, generally leads to superior performance in detecting changes in interval sizes when melodies are shifted in pitch. But the key-distance effect appears qualitatively over a wide range of task difficulty and is even present with kindergarteners, who are not able to solve the transposition detection task at all.

It should be noted that professional musicians might perform flawlessly on our task, showing no key-distance effect due to floor and ceiling effects (although professionals would also be able to verbalize key distance, even if they did not show it in discrimination failures). Still it is interesting that intelligent adults with some degree of musical experience find transposition recognition difficult and are especially prone to confusion between transpositions and same-contour comparisons in a similar key. In this type of short-term memory task, sensitivity to key distance seems easier and more natural than sensitivity to melodic-interval size.

The observed invariance of the key-distance effect strongly supports the psychological reality of key and mode in music memory (Dowling, 1978). At the same time, this invariance across such a wide range of age and experience raises the possibility that key-distance effects result from the processing of auditory information in memory in ways that have little to do with music. The simplest possible auditory theory attributes the key-distance effect to short-term memory for absolute pitch (Deutsch, 1975). In our experiments key distance is correlated with the number of new pitches in the comparison, which did not occur in the standard (Table 1). For example, in Experiment 1 LS comparisons introduced on the average 2.58 new pitches out of 5 (52%), LN items introduced 3.22, and LF items introduced 4.44. Parallel trends occurred with the lure types in Experiments 2 and 3. Hence, one might explain the key-distance effect in terms of repeated and nonrepeated pitches, that is, the greater the number of new pitches in a comparison stimulus, the more different it sounds, and the lower the probability of a response of *same*. To evaluate this pitch-repetition account, we performed correlations across items between new pitches in the comparison and the false alarms it generated for LS, LN, and LF items grouped together. In Experiments 1 and 4, this correlation was not significant, being close to zero in Experiment 1. However, the correlations for Experiments 2 and 3 did reach signifi-

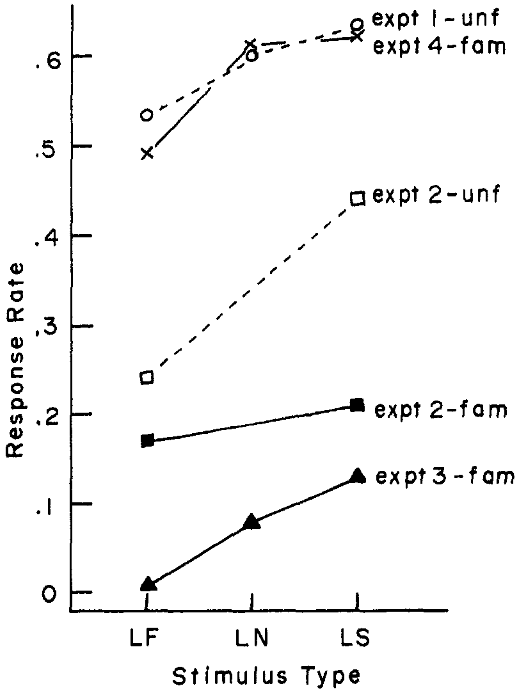

Figure 3. False alarm rates to LF, LN, and LS lures for familiar and unfamiliar standards in Experiments (expt) 1–4. (The data from all experiments are collapsed over experience and age groups. unf = unfamiliar; fam = familiar.)

cance with a one-tailed test at the .05 level, giving $r(18) = -.39$ and $r(13) = -.49$, respectively. The pitch repetition view can neither be accepted nor rejected on this evidence.

We prefer the view that key-distance effects reflect musical schemata that are to some degree culture specific but acquired early in life. According to this view a listener hearing a melody assimilates it to an internal schema that represents a particular mode in his culture. This modal scale is temporarily anchored to a specific pitch level, thus representing what we are calling a key. The mode-key schema governs expectations for the pitches of subsequent notes in the melody. The occurrence of a new pitch is disruptive only if it violates the schematic representation of the key. Thus, it should not be the occurrence of new notes per se that causes an impression of dissimilarity but rather the occurrence of notes foreign to the mode-key schema. However, we have no direct evidence that mode schemata influence performance in short-term memory tasks such as those used here. A critical test of a pitch similarity versus a mode schema account might involve a rigorous control of the number of new pitches that distinguish LN and LF comparisons from standard stimuli. Then differences in false alarm rates to LN and LF comparisons having the same number of new pitches could not be attributed to absolute pitch memory and would support the schema view. Such an experiment remains to be done.

The present results raise another question that is completely independent of the pitch-repetition versus mode-schema issue. This question concerns the relative retention characteristics of key and scale information versus melodic-interval information. It is frequently stated that people do not generally remember the absolute pitches in melodies for appreciable lengths of time but that long-term memory for melodies involves memory for only interval sizes. Dowling (1978) took this view, and Deutsch (1975) stated that

> we recognize a transposed melody much more readily than we recognize the key it was played in. Memory for the abstracted relationships between the component tones of the melody must therefore be more enduring than memory for the tones themselves. (p. 122).

Few would quarrel with these words, yet the present experiments show that with short retention intervals detection of key similarity appears much easier than detection of invariance of melodic-interval size. It would appear that the relative importance of key information in memory falls drastically as retention interval is lengthened, whereas the relative importance of interval information increases. This question, too, awaits further research.

A final issue concerns our failure to observe a significant key-distance effect with targets in Experiments 2, 3, and 4. Ceiling effects on target recognition might explain this failure with familiar melodies and adult subjects in Experiments 2 and 3 (Tables 4 and 5), but target recognition was .75 or lower with unfamiliar melodies and adult subjects (Experiment 2) and with familiar melodies and children (Experiment 4). Yet, a reliable key-distance effect was never obtained with targets. We see three possible explanations for this invariance of target recognition. The first[1] is that subjects might perceive the tonality and key of transpositions more easily than those of lures used in these experiments. Transpositions (as well as standards) often began on the tonic of their key, and, when they did not, the tonic was usually the first accented note of the melody. This was less often true of LS, LN, and LF items, and it is plausible that subjects had more difficulty extracting information about key from lures than from targets. It is possible that a key-distance effect occurs only when there is successful extraction of key information from a standard stimulus and failure to extract key information from a comparison. In such cases, subjects might be influenced by the compatibility between the key of the standard and the notes of the comparison stimulus. This compatibility between key of standard and notes of comparison would be stronger for LS and LN items than for LF. In con-

[1] We thank an anonymous reviewer for suggesting this alternative.

trast, when subjects are able to extract key information from both standard and comparison stimuli, they might be able to separate key information from interval information and to base their decisions solely on the latter. A second possible explanation for the absence of a key-distance effect with targets is the relatively small range of key distance sampled with these stimuli. Whereas the key distance of lures varied from same-key to far-key, that of targets varied only from near-key to far-key. (As mentioned previously, music theory does not allow the possibility of a transposition in the same key as a standard.) A third explanation is that a large shift in key between a standard and a transposition makes the interval matches more noticeable to the listener. A near-key transposition will certainly sound similar to its standard, but listeners might be unable to analyze the source of this similarity. Knowing that their task is to recognize transpositions only, they may often reject such near-key items because they are unsure that interval matches are contributing to the similarity they experience. A far-key transposition might be less confusing in this respect. Listeners might be aware of the changed key (or changed notes), so that any similarity they detect can be attributed to interval matches. Hence, the greater noticeability of interval matches with far-key transpositions might mask the key-distance effect with these items. The present research has clearly raised more questions than it has answered, and the most puzzling of these concern the mechanism by which subjects separate interval matches from other sources of similarity between melodies.

References

Attneave, F., & Olson, R. K. Pitch as medium: A new approach to psychophysical scaling. *American Journal of Psychology*, 1971, *84*, 147-166.

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

Deutsch, D. The organization of short-term memory for a single acoustic attribute. In J. A. Deutsch (Ed.), *Short-term memory*. New York: Academic Press, 1975.

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

Dowling, W. J., & Fujitani, D. S. Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, 1971, *49*, 524-531.

Francès, R. *La perception de la musique*. Paris: Vrin, 1958.

Grier, J. B. Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, 1971, *75*, 424-429.

Imberty, M. *L'acquisition des structures tonales chez l'enfant*. Paris: Klincksieck, 1969.

Riley, D. A., McKee, J. P., Bell, D. D., & Schwartz, C. R. Auditory discrimination in children: The effect of relative and absolute instructions on retention and transfer. *Journal of Experimental Psychology*, 1967, *73*, 581-588.

Swets, J. A. The relative operating characteristic in psychology. *Science*, 1973, *182*, 990-1000.

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

Zenatti, A. Le développement génétique de la perception musicale. *Monographies Francaises de Psychologie*, 1969 (Whole No. 17), 1-110.

Received June 25, 1979 ■