Music Perception
Winter 2002, Vol. 19, No. 2, 249–276

Memory and the Experience of Hearing Music

W. JAY DOWLING
University of Texas at Dallas

BARBARA TILLMANN
Dartmouth College

DAN F. AYERS
University of Texas at Dallas

We report five experiments in which listeners heard the beginnings of classical minuets (or similar dances). The phrase in either measures 1-2 or measures 3-4 was selected as a target, tested at the end of the excerpt. A "beep" indicated the test item, which was a continuation of the minuet as written. Test items were *targets* (repetitions of the selected phrase), *similar lures* (imitations of targets), or *different lures*, and occurred after delays of 4–5, 15, or 30 s. We estimated the proportion of correct discriminations of targets from similar lures and targets from different lures. In Experiment 1, discrimination of targets from similar lures (but not of targets from different lures) improved between 5 and 15 s. Experiment 2 extended this result to a delay of 30 s. Discrimination of targets from similar lures improved over time, especially for second-phrase targets. This improvement was due mainly to decreasing false alarms to similar lures. Experiments 3 and 4 replaced the continuous music with silence and with a repetitive "oom-pah-pah" pattern, and the improvement in discrimination of targets from similar lures disappeared. Experiment 5 removed listeners' expectations of being tested, and the improvement also disappeared. Results are considered in the framework of current theories of memory, and their implications for the listener's experience of hearing music are discussed.

Received June 18, 2001, accepted July 26, 2001.

MEMORY is always changing; sometimes it improves; sometimes it gets worse; rarely does it stay the same. The changes in memory over time are qualitative as well as quantitative. At times one cue will be most

Address correspondence to W. Jay Dowling, Program in Cognitive Science, University of Texas at Dallas, Richardson, TX 75083-0688. (e-mail: jdowling@utdallas.edu)

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

effective in evoking a particular memory; at times another will be most effective. That is, what we remember of an event changes; in its early trajectory, it is characterized by one set of features, and later it is characterized by others.

Considerable converging evidence shows quantitative and qualitative changes in the listener's memory for a musical phrase during the first few minutes after its initial hearing. We review the evidence from recent studies and suggest theoretical approaches that offer the promise of providing a basis for explaining the phenomena. We then present five new experiments that increase ecological validity over previous studies and that test aspects of those theories.

Changes in Melody Recognition Over Time

Dowling and Bartlett (1981) and DeWitt and Crowder (1986) had observed changes in the effectiveness of various retrieval cues over time in melody recognition. In a study focusing on that phenomenon, Dowling, Kwak, and Andrews (1995, Experiments 1–3) tested recognition of brief, novel, isochronous melodies after different delays. To assess the relative contributions of various melodic features to recognition performance, they used three types of test items: targets (T, exact repetitions of earlier melodies), similar lures (S, same-contour imitations of targets), and different lures (D, new melodies with different contours). These items were presented intermixed as a continuous series of melodies. Listeners responded to each melody presented, judging whether it was the same as, or different from, previous melodies in the list (a continuous-running-memory task; Shepard & Teghtsoonian, 1961). Memory for a given melody was tested after a time interval filled with other items and responses. The delay between the presentation of a new item and its test varied systematically. Listeners' discrimination between T and D items (T/D discrimination) declined over time, although T/S discrimination did not decline over delays of up to 1.5 min. From this result, Dowling et al. inferred that the relevance of the various melodic features to recognition changes over time.

Dowling et al. (1995, Experiments 5–7) moved in the direction of greater ecological validity by applying the same paradigm to the melodic lines of unfamiliar folk songs. With these richer materials they found that T/S discrimination actually improved across filled delays up to 2 min, with proportion correct (estimated by area under the memory-operating-characteristic curve; Swets, 1973) going from 0.65 at 22 s to 0.70 at 2 min. This improvement was largely due to a decrease in false-alarm rates to S items; that is, listeners found it easier to reject similar lures after the longer delay. During the same period, T/D discrimination declined from 0.76 to 0.73.

Again, there was a change over time in the relevance of various melodic features to recognition. And the shift to more natural melodies led to a clear differentiation between improvement for the finer T/S discrimination and decline for T/D discrimination.

Dowling, Barbey, and Adams (1999) explored longer delays using the same paradigm. They found that memory for folk song melodies showed slight improvements in both T/S and T/D performance up to about 5 min and a definite decline by 11 min (Experiments 1 & 2). (Note that this is the only instance in either the previous or the present experiments in which T/D performance improved as much as T/S discrimination improved.) We believe it is safe to say that the improvement in melody recognition that concerns us here occurs during the first 5 min after the introduction of a novel melody.

In all the studies just reviewed, the time delay over which listeners remembered novel melodies varied directly with the number of intervening trials between a new melody and its test. To separate the effects of time delay and intervening trials, Dowling et al. (1999, Experiment 3) introduced two or four intervening trials during delays of 1.5 or 3.0 min in a 2 x 2 design. Overall, T/S discrimination improved from 0.55 to 0.62 between 1.5 and 3.0 min. With two intervening trials, T/S performance improved dramatically over time, going from 0.48 to 0.68. When the number of intervening trials was proportional to time (with two trials filling the 1.5-min delay and four filling the 3.0-min delay), a more modest but significant improvement (48% to 56%) was observed.

In summary, previous research using the continuous-running-memory task with natural melodies demonstrates an improvement in T/S discrimination across a filled delay. Erdelyi (1996), in his review of hypermnesia (spontaneous memory improvement) over the past century, notes that such effects in recognition (as opposed to recall) are quite rare. Erdelyi's review suggests that hypermnesia is found more often with complex, highly structured materials such as poetry than with the less structured materials of the American verbal learning tradition. Erdelyi notes that two additional conditions favor the appearance of hypermnesic effects: first, continued presentation of competing material during the delay between the initial presentation of an item and its test; and second, the subject's expectation of being tested. Both of these conditions characterize the studies of Dowling et al. (1995, 1999) just reviewed. We test the possible importance of the expectation of a test in Experiment 5.

Theoretical Approaches

Confronted with these surprising improvements in T/S discrimination over time in recognition memory, we sought a theoretical approach to hu-

man memory that would suggest possible explanations for the effects observed. Most contemporary theories of memory do not provide for improvement over time, and especially do not consider the possibility in recognition. For most memory models, the strength of a memory trace naturally declines over time or is interfered with by other material. Enhancement over time is not usually provided for. It is difficult for us to see, for example, how the family of models related to Gillund and Shiffrin's (1984; cf. Estes, 1999) SAM model could explain recognition improvement over time, though we cannot categorically rule out improvement in recall with such a model, because recall involves complex processes of memory search. There are two theoretical approaches that we think might well be applicable to our results: (1) Tulving's theory of episodic memory and (2) his considerations of a retrieval strategy shift between short-term and long-term memory.

Tulving's (1983, 1984/1986, 1985) theory of episodic memory proposes that memory stores traces of past events via a process of encoding. When a retrieval cue is presented, the memory is queried. Information from the cue is combined with information from the retrieved trace to produce a set of "ecphoric" information. This process of recoding with the ecphoric information can change the original trace in memory. The ecphoric information—not necessarily the same as that of the original trace—provides the basis for the memory system's response to the query.

An example of the kind of phenomenon Tulving's theory was developed to explain—a parallel in the verbal domain to the results in music just reviewed—is provided by Tulving and Watkins (1975). In a study of cued recall, Tulving and Watkins found changes over time in the effectiveness of various retrieval cues. Subjects learned word lists under different encoding conditions. Target words were initially paired with rhyming cues (for example, "chair" with "pair") or with semantically associated cues (for example, "chair" with "desk"). Then Tulving and Watkins assessed the effectiveness of new rhyming ("hare") or associative ("table") cues in evincing recall of the targets during two tests separated by other test trials (a matter of minutes). These new cues provide different features of the original cues, just as in the present experiments the recognition cues T and S provide different sets of features of the targets. In their data analysis, Tulving and Watkins were able to obtain separate estimates of the effectiveness of the various cue features in bringing about recall. Most relevant to our study, they found that whereas some cue elements declined in effectiveness over time, others gained (see their Table 6, p. 272). Tulving's theory of episodic memory is designed to explain the implications of this kind of result; namely, changes over time in the information available in a memory trace.

There are three aspects of Tulving's approach that strike us as offering the beginnings of an explanation for the results of Dowling et al. (1995,

1999) and those of the present experiments. First, it provides an account of changes over time in the information available in memory. Second, these changes in trace information lead to changes in the effectiveness of various retrieval cues, as seen in Tulving and Watkins (1975) and Dowling et al. (1995, 1999). Third, one source of information contributing to recoding comes from test items presented during the delay as in the continuous-running-memory task (Dowling et al., 1995, 1999). We believe that it is possible that recoding might occur even without explicit recall, just as the result of continued attentive listening in the delay between the presentation of a target and its test, as in the present experiments.

Two recent formal models of distributed memory seem to us to capture aspects of Tulving's general theory and offer the promise of providing a precise characterization of results such as those of the present experiments; namely, Murdock's (1997, 1999) TODAM model (in its chunking process), and Goldinger's (1998) adaptation of Hintzman's MINERVA model (with its echoing process). Both incorporate a version of recoding, and that is the essential feature that offers promise for explaining our results in music.

The second approach that offers the possibility of explaining our results was perhaps most clearly stated by Tulving (1987) in his discussion of an experiment by Wright, Santiago, Sands, Kendrick, and Cook (1985), though a version of the theory was proposed by Dowling and Bartlett (1981) to explain some precursors of the results of Dowling et al. (1995, 1999). Wright et al. gave the same task to pigeons, monkeys, and people. They presented a series of four pictures, followed by either an immediate or a delayed recognition test. The test item was one of the items presented or a different lure. Wright et al. found a strong recency effect with the immediate test; that is, poor memory for early items in the list, and good memory for the last ones. However, after longer (unfilled) delays (up to 100 s for people) the pattern shifted to a primacy effect, with the early list items better remembered than the later. For example, for people the first list item went from about 67% correct on immediate test to about 90% after 100 s. Tulving (1987) suggests that the recency effect on the immediate test could conceivably arise from the operation of the short-term memory system, such that "the information in short-term memory inhibits the use of related information in long-term memory" (p. 71). With longer delay, greater weight is given to the output of the long-term memory system, and the release from inhibition is seen in the improvement of performance with early items.

The biggest difference between these two approaches would seem to be their predictions in cases in which the delay interval between the presentation of an item and its test is filled versus empty. A delay filled with continued presentation of music would seem to be necessary for recognition improvement to occur according to the recoding hypothesis, whereas the shift between short-term and long-term retrieval systems should operate even

across empty delays (as in Wright et al., 1985). We test some of these implications in the experiments presented next.

New Experiments on Memory Improvement Over Time in Music

Here we present a series of new experiments that explore the phenomena of changes in memory for musical phrases during the first minute after they are heard. The sequence of experiments just reviewed (Dowling et al., 1995, 1999) moved from artificial materials (isochronous melodies) in the direction of more and more "natural" music, with the result that changes in what is remembered began to appear as an improvement in T/S discrimination over time. Here we continue that progress toward greater ecological validity by using original minuets from the repertoire. This type of ecological material has been used recently in target detection and recognition tasks studying the perception of global organizational structures in music (Tillmann & Bigand, 1998).

In the present case, it occurred to us that the patterns of repetition and imitation in compositions (such as minuets) in the Western European tradition (as in other traditions, Dowling & Harwood, 1986) provide an opportunity to test memory for musical phrases in the context of more or less "natural" music listening. Composers vary the time delay between the presentation of a motif and its repetition or imitation; compare the opening movement of Beethoven's Fifth Symphony, in which the initial motif is imitated immediately and over and over, and the opening of Brahms's Second Piano Concerto, in which the opening measure in the horn is echoed only some 20 measures later by the full orchestra, and by the piano soloist only after another 42 measures. We realized that we could use this "natural" variation in delay of repetition and imitation to create tests involving T (repetition) and S (imitation) items, as well as D (different) items, after various delays. We turned to classical minuets as a source of stimuli because they afford testing at relatively brief delays and because there are a large number of them available that are consistent in style and form.

Consider Figure 1A, the well-known Minuet in G by Beethoven. The intial phrase (bracket 1) is followed by a different second phrase (bracket 2), but then imitated in the third phrase (bracket 3). We could use that third phrase as an S test item for the first phrase, after a short delay (5 s). Or we could wait longer for a test of the second phrase. Note that with the repeat the second phrase (bracket 2) is repeated exactly and that there are no intervening repetitions or imitations of that second phrase. Thus the second phrase in the repeat provides a T item at a longer delay (15 s). D test items can be found wherever new material is introduced, as at bracket 7.

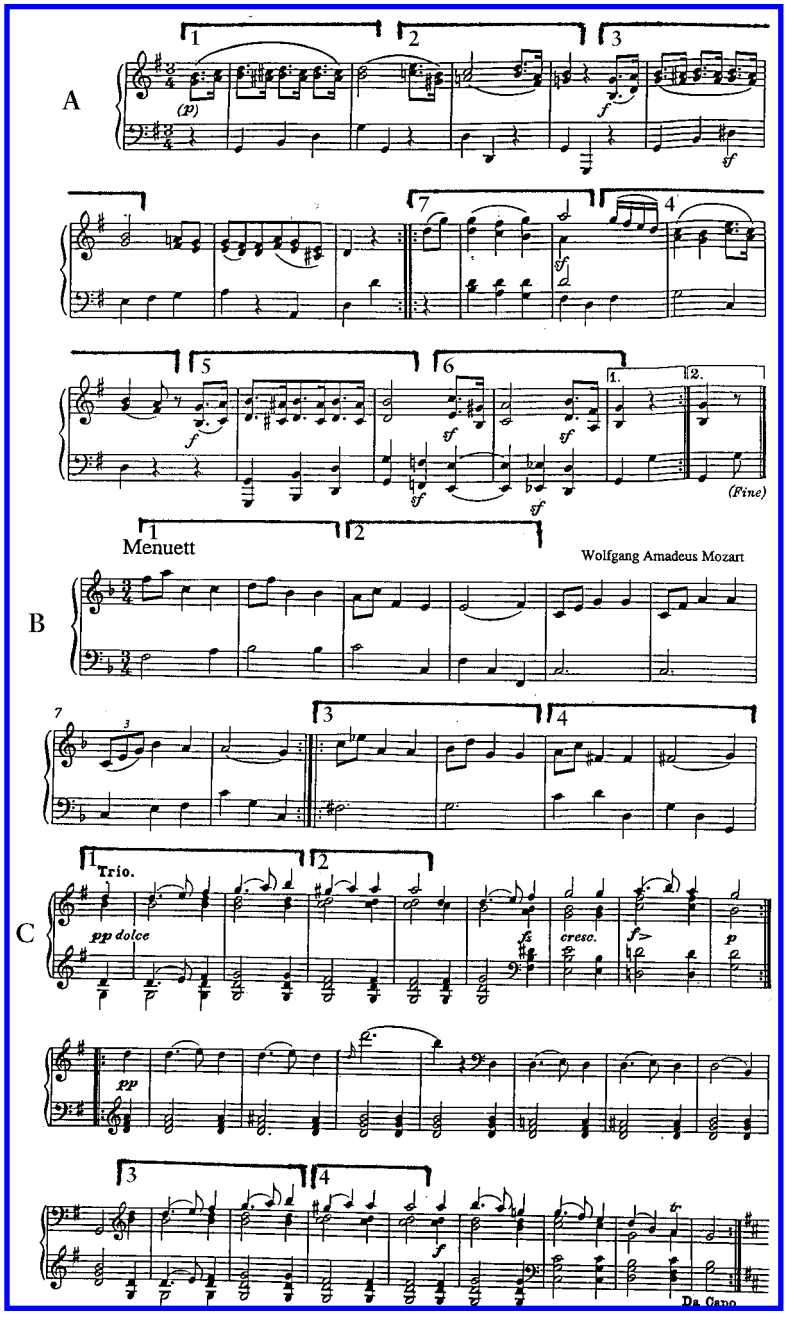

Fig. 1. Examples of stimuli. The brackets indicate placement of possible target (1 and 2) and test (3 and 4) phrases. (A) Beethoven, Minuet in G. Bracket 5 indicates a potential test phrase that cannot be used because of the imitation at bracket 3. The test at bracket 6, however, is possible. (B) Mozart, Minuet in F, K. 2. (C) Schubert, Waltz, Op. 127, No. 1, Trio, D. 146.

Note that these sequences of presentation, delay, and test follow exactly what Beethoven wrote.

Another example is provided by a minuet that Mozart wrote when he was six (Fig. 1B). The initial phrase (bracket 1) could be tested after 15 s (6 measures) as either T (bracket 1, with the repeat) or S (bracket 3). Similarly the second phrase (bracket 2) could be tested at 15 s as T (bracket 2) or S (bracket 4).

In each piece, we selected either the phrase in measures 1-2 ("first phrase") or the phrase in measures 3-4 ("second phrase") as the target. A soft, high-pitched "beep" called attention to the start of the test phrase.

Tests at 30 s generally made use of a recapitulation midway through the second section of the piece. In Figure 1A such a test of the second-phrase target (bracket 2) could occur with an S comparison at bracket 6. An S test of the first-phrase target (bracket 1) at bracket 5, however, would be disqualified by the earlier imitation at bracket 3. The Schubert waltz in Figure 1C provides an example in which either the first or second phrase (brackets 1 and 2) could be tested as T after a 30-s delay of 14 measures (brackets 3 and 4).

In Experiment 1, we applied this new method using delays of 5 and 15 s between the presentation of a phrase and its test. In Experiment 2, we extended the delays to 30 s. There we also constructed counterbalanced lists in order to avoid the possibility that our observed effects might be due to fortuitous combinations of particular minuets and particular conditions. Because interference effects in memory are well known, Experiments 3 and 4 addressed the hypothesis that removing the musical material between the introduction of an item and its test could lead to even stronger improvement in performance. In Experiment 3, we omitted the intervening material. In Experiment 4, the intervening material was replaced with a metrical but musically meaningless "oom-pah-pah" pattern. As noted above, the manipulations of Experiments 3 and 4 bear on the differentiation of the recoding and short-term vs. long-term retrieval strategies hypotheses. And because Erdelyi's (1996) review suggested that the listener's expectation of being tested might be an important determinant of performance, in Experiment 5 we led listeners to believe that there would be no test.

Experiment 1

METHOD

Listeners

Undergraduates at the University of Texas at Dallas (mean age, 25.2 years) participated in the study as part of their course requirements in psychology. Those categorized as moderately experienced had at least 2 years of explicit musical training (defined as lessons on an

instrument or voice, or playing in an instrumental ensemble; mean = 6.3 years, SD = 3.6 years). Those with less training were categorized as musically inexperienced. Twenty-two listeners served in Experiment 1, of whom 10 were moderately experienced and 12 inexperienced.

Stimuli

The stimuli were drawn from classical minuets, waltzes, and German dances or Ländler written for piano between 1750 and 1828 by Haydn (1984, 1989), Mozart (1956, 1992), Beethoven (1967, 1987, 1990) and Schubert (1989—see also Lakos, 1994). (We included some minuets from Mozart's sonatas for violin and piano, adding the violin notes to the piano part.) These dances in 3/4 time followed a form in which an initial section of 8–12 measures (delimited by a repeat sign) was followed by a (usually longer) section of 8–32 measures (see Figure 1).

On each trial, listeners heard the first 15 to 25 s of a piece. During the first 10 s, a phrase was chosen as a target to be tested later. This target was either the first phrase (approximately measures 1 and 2) or the second phrase (approximately measures 3 and 4) of the piece. A new piece appeared on every trial. Pieces were selected so that the target phrase was neither repeated nor imitated in the continuation of the piece prior to the test. Test phrases occurred within the piece after delays of 5 or 15 s—2 or 6 measures at a tempo of 72 beat/min.

The music continued just as written by the composer following the presentation of the target. The onset of the test was signaled by a soft, high-pitched "beep" that did not interfere with the music, occurring one-half beat before the test. The music stopped after presentation of the test phrase, and listeners were given 10 s to respond before the beginning of the next trial.

The test phrase was either an exact repetition of the target (T) or an imitation of the target that changed one or more features (S), or a totally new phrase not heard before in the piece (D). S phrases shared the melodic contour (overall pitch and rhythmic pattern) of targets, but differed in pitch level, texture (number of simultaneous voices and their ranges and density), accompanying chords or rhythms, or some combination of those. D phrases differed from targets at least in melodic contour and usually in several features.

There were 12 types of trial defined by the combination of two types of target (first or second phrase), two delays (5 or 15 s), and three types of test phrase (T, S, or D). There were 48 trials in all, with four of each type. The order of trials was randomized so that the list consisted of four permutations of the 12 trial types, ensuring that each type of trial would be tested equally often toward the beginning and toward the end of the list. The list of trials was divided into two equal sections and half the listeners heard the sections in reverse order.

Stimuli were played on a Yamaha Clavinova P-100 (which has weighted piano keys and touch-sensitive response) and recorded by a PC-type computer via its MIDI interface. Particular attention was given to articulation, phrasing, and dynamics to make the performance as natural and aesthetically pleasing as possible. We used Cakewalk software to edit the recordings, correcting recording errors, and imposing a uniform tempo. Tempos were selected to produce the appropriate delay between the introduction of an item and its test. Even though local tempo variations within each piece would have sounded more natural, such variations are context dependent—determined by the place of a phrase within a section (Gabrielsson, 1999). That would have led to tempo gradient differences between targets and test items, distinguishing test items from the rest of the piece. Therefore we used a uniform tempo within each piece. We used the Cakewalk editor to insert exact repetitions to serve as T test items. Stimuli were played for listeners on a Yamaha TG-500 synthesizer using its "acoustic piano" voice, computer-controlled via MIDI interface, and presented to listeners in group sessions via loudspeakers at comfortable levels.

Procedure

Listeners were introduced to the experiment by a brief explanation of the task including examples of the differences among T, S, and D test items. The instructions emphasized that

listeners should respond "same" only when the test item was exactly the same as the target, and they should reject S as well as D lures. (Earlier work indicates that listeners would indeed find it difficult to do otherwise; Dowling & Bartlett, 1981.) Listeners were instructed to respond using a six-point confidence-level scale on which 6 meant "very sure same," 5 "sure same," 4 "same," 3 "different," 2 "sure different, and 1 "very sure different." Listeners also completed a brief questionnaire concerning musical experience.

Data Analysis

The data from Experiment 1 were analyzed in a 2 Experience Levels × 2 Test Delays × 3 Test Items (T, S, D) design. All but the first of those variables involved within-groups comparisons. Responses to the three test item types were reduced to two areas under the memory-operating-characteristic curve, one assessing discrimination between T targets and S lures (T/S), and the other discrimination between T targets and D lures (T/D).

The six-point scale provided us with five criterion placements on the memory-operating-characteristic curve with which to calculate the area. Area under the memory-operating-characteristic curve provides an unbiased estimate of performance where chance is 0.50 (Swets, 1973). The area score provides a better measure of performance than, for example, d' based on the criterion between responses 3 and 4, because it preserves more response information and over the years has proved to be uncorrelated with measures of bias (unlike d'—see, for example, Dowling et al., 1995).

We report analyses of area scores and of individual listeners' median ratings of the stimulus types on the six-point scale. We chose to use the median ratings as an index of listeners' tendencies to respond positively to Ts (hits) and to Ss and Ds (false alarms) for two reasons. First, the ratings provide a more sensitive scale than proportions of 4-5-6 responses (contrasted with 1-2-3 responses). Second, using the median rating for each listener for each condition lessens the effect of specific items that might be outliers in the distribution of responses to a particular item type. We shall refer to the means of these median ratings as "hit ratings" when they pertain to T trials, and as "false-alarm ratings" on S and D trials.

RESULTS

Area scores were subjected to a 2 Experience Levels × 2 Target Positions × 2 Time Delays × 2 Item Types analysis of variance (ANOVA). Since the Delay × Item interaction is crucial to our argument, we shall report it first for each ANOVA, so that it will be easy to find. The Delay × Item interaction was significant, $F(1,20) = 10.05$, $p < .01$, $R^2 = 0.026$ (Table 1). Performance increased substantially with time for T/S comparisons, although it increased only a little for T/D comparisons. The effect of target position was significant, $F(1,20) = 45.76$, $p < .01$, $R^2 = 0.172$, with performance better for first-phrase targets (0.84) than for second-phrase targets (0.66). The effect of delay was significant, $F(1,20) = 6.63$, $p < .02$, $R^2 = 0.051$, with better performance at 15 s than at 5 s (0.80 vs. 0.70). The effect of item was significant, $F(1,20) = 18.47$, $p < .01$, $R^2 = 0.047$, with better performance on T/D discrimination (0.79) than T/S discrimination (0.70). No other effects or interactions were significant.

We calculated median ratings of each stimulus type at each delay for each listener and subjected them to a 2 Experience Levels × 2 Target Positions × 2 Time Delays × 3 Items ANOVA. The Delay × Item interaction

TABLE 1
Area Under the Memory-Operating-Characteristic Curve for T/S and T/D Comparisons at Two or Three Time Delays in Experiments 1-4

	Comparison					
	T/S			T/D		
Experiment	5-s delay	15-s delay	30-s delay	5-s delay	15-s delay	30-s delay
1: music	0.61	0.78		0.78	0.81	
2: music	0.66	0.72	0.72	0.79	0.77	0.77
3: empty delay	0.87	0.86		0.97	0.95	
4: oom-pah-pah	0.85	0.84		0.95	0.93	

T/S indicates comparisons of targets and similar lures; T/D indicates comparisons between targets and different lures. The shortest delay in Experiment 2 was 4 s.

$F(2,40) = 5.82$, $p < .01$, $R^2 = 0.028$, was significant. False alarms to S lures declined sharply over time, whereas hits and D false alarms remained about the same (Table 2). Significant main effects such as those of target position and delay that do not involve item type merely reflect some mixture of shifts in performance and in response criteria. The effect of target position was significant, $F(1,20) = 9.59$, $p < .01$, $R^2 = 0.016$, as was the effect of delay, $F(1,20) = 31.52$, $p < .01$, $R^2 = 0.026$. The effect of item was significant, $F(2,40) = 63.44$, $p < .01$, $R^2 = 0.342$, indicating that listeners discriminated among the items (means of 4.1, 2.8, and 2.0 for median ratings of T, S, and D items, respectively). The Position × Item interaction was significant, $F(2,40) = 25.66$, $p < .01$, $R^2 = 0.051$. Second-phrase targets produced lower hit ratings and higher false-alarm ratings than did first-phrase targets; that is, they led to poorer discrimination. No other effects or interactions were significant.

TABLE 2
Means of Listeners' Median Ratings of T, S, and D Stimuli at Two Time Delays in Experiments 1–4

	Delay (s)								
	Item T			Item S			Item D		
Experiment	5	15	30	5	15	30	5	15	30
1: music	4.3	4.0		3.4	2.3		2.1	2.0	
2: music	4.4	4.2	4.0	3.2	2.8	2.6	2.3	2.1	2.4
3: empty	5.1	4.5		2.2	1.8		1.2	1.1	
4: oom-pah-pah	4.9	4.6		2.3	2.1		1.2	1.2	

T indicates targets, S indicates similar lures, and D indicates different lures. The shortest delay in Experiment 2 was 4 s.

DISCUSSION

Just as with the continuous-running-memory paradigm, performance improved over time for the T/S discrimination, in this case by 17% (Table 1). And as before, this improvement was due principally to a decline in false alarms to S lures (Table 2).

Experiment 1 only tested recognition at delays up to 15 s. Since previous studies (Dowling et al., 1995, 1999) had shown improvement at longer delays, we wanted to see if that would be the case with the new paradigm. Therefore in Experiment 2, we added a 30-s delay (similar to that shown in Figure 1C).

Because with stimuli consisting of actual music there is typically a large amount of variability in the responses attributable to particular pieces, we also counterbalanced Experiment 2 so that a piece tested at one delay for one group of listeners would be tested at another delay for another group. This counterbalancing was made possible by having three delays. That is, it was not always possible to use a given piece in exactly the two delays of Experiment 1, whereas, given an initial assignment to a delay condition, it was usually possible to shift the piece to one of the other two remaining delays in Experiment 2.

Experiment 2

METHOD

Fifty-two listeners served in Experiment 2, of whom 30 were moderately experienced and 22 inexperienced. Experiment 2 was like Experiment 1, except that there were 72 trials with tests at delays of 4, 15, and 30 s, giving four instances each of the 18 types of trial. We shifted to 4 s for the shortest delay to make tempi a little faster (90 beats/min), producing a better match on the average with the tempos of stimuli at the two longer delays, especially the 30-s delay. (That is, the 30-s delay did not always comprise 12 measures, so the tempo was adjusted to preserve the exact time delay.) A pilot study that replicated Experiment 1 at 90 beats/min with delays of 4 and 12 s (90 beat/min) convinced us that the change in tempo had little effect on performance, but did make the pieces more interesting to listen to.

We constructed counterbalanced lists for Experiment 2 such that almost all items tested at a given time delay were tested at another time delay in the other list. This imposed additional constraints on the selection of stimuli. In all but two cases, both of them involving the 30-s delay, we were able to find pieces that could be tested at two delays, typically with different target positions and item types. Roughly equal numbers of pieces were tested in each of the possible pairs of delays: 26 in 4 and 15 s; 23 in 4 and 30 s, and 21 in 15 and 30 s. The two pieces that appeared twice at the 30-s delay shifted between target positions and item types.

RESULTS

Area scores were subjected to a 2 Experience Levels × 2 Counterbalanced Lists × 2 Target Positions × 3 Time Delays × 2 Items ANOVA. The

Delay × Item interaction was significant, $F(2,96) = 6.24$, $p < .01$, $R^2 = 0.008$ (Table 2). Performance increased over time for T/S comparisons, at least from 4 to 15 s, whereas performance remained unchanged for T/D comparisons. The effect of target position was significant, $F(1,48) = 24.44$, $p < .01$, $R^2 = 0.053$, with performance better for first-phrase targets (0.78) than for second-phrase targets (0.69). The effect of item type was significant, $F(1,48) = 56.71$, $p<.01$, $R^2 = 0.041$, with better performance on the T/D comparison (.78) than T/S (.70). The Position × Delay interaction was significant, $F(2,96) = 4.19$, $p < .02$, $R^2 = 0.013$, in which performance improved over time for second- but not for first-phrase targets. The Position × Item interaction was significant, $F(1,48) = 37.49$, $p < .01$, $R^2 = 0.016$. T/D performance decreased much more in going from first- to second-phrase targets (0.85 to 0.71) than did T/S performance (0.72 to 0.68). Finally, the interaction of Position × Delay × Item × List was significant, $F(2,96) = 5.84$, $p < .01$, $R^2 = 0.007$. T/S performance generally improved with time, except for first-phrase targets on List 2. T/D performance declined with time for first-phrase targets, but produced mixed performance for second-phrase targets. No other effects were significant.

The median ratings of each stimulus type at each delay for each listener were subjected to a 2 Experience Levels × 2 Counterbalanced Lists × 2 Target Positions × 3 Time Delays × 3 Item Types ANOVA. The Delay x Item interaction, $F(4,192) = 3.54$, $p < .01$, $R^2 = 0.007$, was significant (see Table 1). Ratings of Ts decreased, and ratings of Ss decreased even more, whereas ratings of Ds did not change much over time. The Position × Delay × Item interaction, $F(4,192) = 7.09$, $p < .01$, $R^2 = 0.010$, was also significant (Figure 2). Ratings of S lures decreased over time for both first- and second-phrase targets, and moreso for the latter. Ratings of Ts decreased over time for first-phrase targets but remained high for second-phrase targets. D false alarms did not change much. This pattern gives the details underlying the significant interaction of Position × Delay for area scores, and will be discussed later.

As before, effects that do not involve item type merely reflect some mixture of shifts in performance and in response criteria. The effects of target position, $F(1,48) = 25.83$, $p < .01$, $R^2 = 0.002$, delay, $F(2,96) = 7.43$, $p < .01$, $R^2 = 0.009$, and the interaction of Position × Delay × List, $F(2,96) = 7.69$, $p < .01$, $R^2 = 0.006$, were all significant. The effect of item was significant, $F(2,96) = 188.89$, $p < .01$, $R^2 = 0.281$, indicating that listeners discriminated among the items (means of 4.2, 2.9, and 2.3 for ratings of T, S, and D items, respectively). The Position × Item interaction was significant, $F(2,96) = 33.59$, $p < .01$, $R^2 = 0.026$, the main difference being that D lures were easier to reject for first-phrase targets than for second-phrase targets.

The interaction of Position × Item × List was significant, $F(2,96) = 6.60$, $p < .01$, $R^2 = 0.005$. The only departures greater than 0.2 rating points

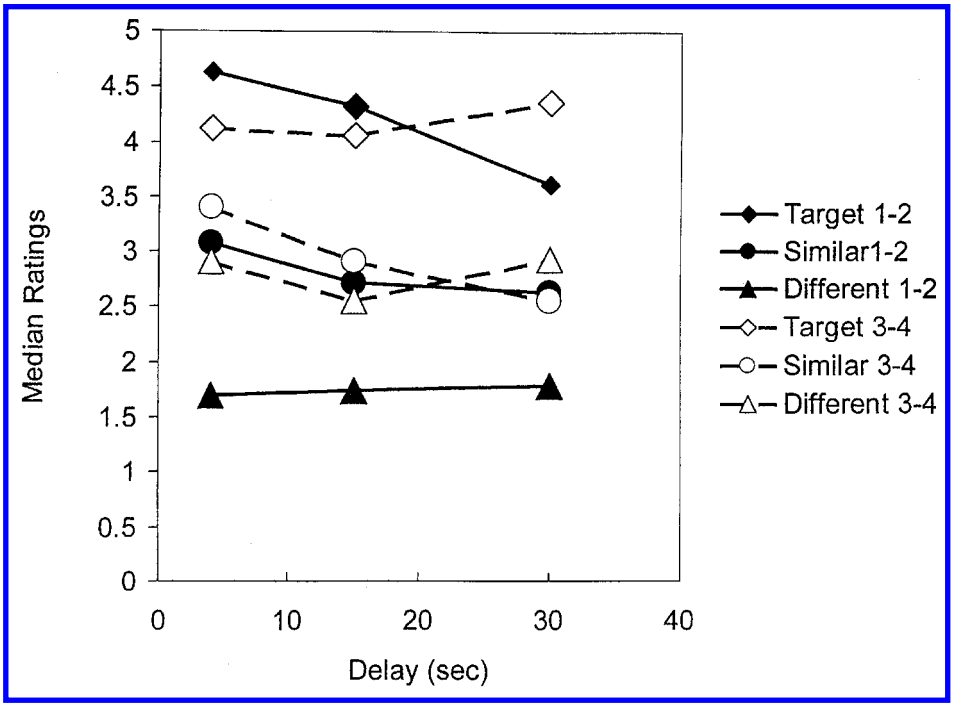

Fig. 2. Means of listeners' median ratings on a six-point scale of target (T), similar lure (S), and different lure (D) test items for first-phrase (1-2) and second-phrase (3-4) targets at three delays in Experiment 2, showing the three-way interaction of those variables. Higher ratings of T indicate hits, whereas higher ratings of S and D indicate false alarms.

from that of the Position × Item interaction is that first-phrase Ts were rated lower on List 1 than on List 2. Finally, the interaction of Position × Delay × Item × List was significant, $F(4,192) = 3.96$, $p < .01$, $R^2 = 0.006$. The pattern shown in Figure 2 holds, except that S ratings for first-phrase targets do not decline over time for List 2, and there is a complex variation in D ratings that can be attributed to the difference between lists. In particular, the decline of S false alarms for second-phrase targets is confirmed. There were no other significant effects.

DISCUSSION

As in Experiment 1, T/S discrimination improved over a filled delay interval. This was especially true for second-phrase targets, with performance going from 0.61 to 0.69 to 0.74. This improvement was characterized by hit rates that remained steady or increased over time, coupled with decreasing S false-alarm rates (Figure 2). The decline in S false alarms suggests that listeners become better at making fine discriminations involving

details of the stimuli. They are able to reject test items that resemble targets in broad outline, but that differ in detail.

The overall decline in S false alarms was more pronounced for second-phrase targets (0.9 rating points) than for first-phrase targets (0.5 rating points). It is apparent from Figure 2 that the interaction involving target position would have been unlikely to occur in Experiment 1. The steadiness of hits and the decline in S false alarms is closely parallel for first-phrase and second-phrase targets between 4 s and 15 s, as in Experiment 1. It is only between 15 and 30 s that the patterns diverge, with hits remaining high and S false alarms continuing to decline for second-phrase targets.

Why should there be more improvement with second-phrase targets? One possibility is that the processes that lead to improvement are automatic, and are easily interfered with by conscious control. First-phrase targets are more likely to be consciously registered as potential test items, and hence more likely to be subjected to controlled processes such as rehearsal during the delay interval. This tendency toward controlled processing leads to superior performance overall (0.78 vs. 0.69, a classic serial-position effect), but does not lead to as much improvement over time. It is the second-phrase targets, less likely to be consciously rehearsed but apparently subjected to other forms of continued automatic processing, for which T/S discrimination improves more strongly.

The use of three delay intervals provided the flexibility to counterbalance particular items across delays. This gives us confidence in Experiment 2 that the improvement observed over time was not simply due to fortuitous pairings of delays with pieces of music. With the counterbalanced lists, we can see the degree to which the pairing of particular conditions and individual pieces affects the outcome. Although most of the ratings show a close convergence between the two lists, there are still puzzling divergences. Though none of these would alter our qualitative conclusions, nevertheless the divergences emphasize the importance of using counterbalanced lists as a safeguard against arbitrary pairings of stimuli and conditions that could affect the conclusions. We emphasize that though the significant Position × Delay × Item interaction was complicated by list effects, T/S improvement over time for second-phrase targets was equally strong for the two lists: about 0.13 between 4 and 30 s, and that improvement was largely attributable to a decline in S false alarms.

Having found significant effects of list, we wanted to find out if individual pieces were easier or harder to judge, independent of the condition. To answer this question, we ran an item analysis, finding the mean of the median ratings for each piece of music as it appeared on Lists 1 and 2. We reversed the sense of the scale for false-alarm ratings, so that correct ratings were always at the high end of the scale. The correlation between

average ratings of the pieces on the two lists was close to zero (-0.05). This means that whatever contributed to the variability between the lists was not the global memorability of individual pieces as such. Rather, what was important was the memorability of particular phrases in context.

Experiments 1 and 2 demonstrate an improvement in recognition memory across time intervals filled with meaningful music, especially for second-phrase targets tested for discrimination between highly similar musical imitations (T/S). In trying to determine what was producing this surprising result, we thought it best to start by examining the role of the musical material intervening between the introduction of a target item and its test. The first manipulation we tried was to eliminate the intervening material, leaving a blank time interval of the same duration. Experiment 3 replicates Experiment 1 but with all music except the target and the test eliminated.

The contrast between blank and filled delays leads to different predictions depending on the theoretical characterization of the role of the intervening material in memory. To start with the theoretical approaches outlined in the Introduction, our interpretation of Tulving's (1983, 1984/1986, 1985) theory of episodic memory, with its provision for recoding during a filled interval, should predict a lack of improvement during the empty interval, during which recoding should not occur. If, on the other hand, the improvement in performance on first and second phrases is due to a shift in retrieval strategy from a short-term system to a long-term system (Tulving, 1987), then that improvement should occur as well across an empty interval (as it did in Wright et al., 1985) as across a filled one.

There are two other theoretical approaches we wish to mention that make predictions concerning filled and empty delays. Many contemporary theories of memory (see Baddeley, 1997) see the intervening material as a source of interference with memory for the previously presented target. A blank interval following an item provides the opportunity for the working memory system to rehearse the material just presented, thereby improving its chances of retrieval. Filling the delay with more music should prevent rehearsal and lead to worse performance. If the process that is involved in the improvement observed in Experiments 1 and 2 can also operate more effectively without interference, an even greater improvement should be observed over time in Experiment 3.

Another approach takes the intervening material not as causing interference but as adding meaning and connection to targets and test items. Part of the meaning of musical phrases depends on their relation to the context around them. And meaningfulness in this sense can enhance memory performance. In an analogous example of memory for chess positions, Frey and Adesman (1976) found that moderately expert players' memory for a briefly presented position in the middle of a game was enhanced by presenting the sequence of moves leading up to that position, as contrasted

with just presenting the position out of context. On this account we would expect that more intervening material in the longer delay intervals would lead to better performance (as in Experiments 1 and 2), and that dropping the intervening material would hurt performance. Listeners' comments lend some support to this approach. Several remarked after Experiments 1 and 2 that on the longer trials they had a better understanding of the piece, and this led them to be more confident of their responses.

Experiment 3

METHOD

Twenty-eight listeners served in Experiment 3, of whom 15 were moderately experienced and 13 inexperienced. Experiment 3 was the same as Experiment 1, except that the music between the presentation of the target and the test phrase was omitted, as was the music preceding the target (if any). That is, there was no uncertainty on the listeners' part concerning the identity of the target. Each trial simply started with the presentation of the target phrase and ended with presentation of the test phrase (preceded by the beep). In addition, we constructed a counterbalanced list in which all 5-s tests were moved to 15 s and vice versa. This was possible because without the intervening music there was nothing that tied a particular test phrase to a particular point in time (which had prevented the use of counterbalanced lists in Experiment 1). Approximately half the listeners performed the experiment with each list.

RESULTS

Area scores were subjected to a 2 Experience Levels × 2 Positions (first-phrase vs. second-phrase) × 2 Time Delays × 2 Items ANOVA. The Delay × Item interaction was not significant; those data are shown in Table 1 for comparison with the other experiments. Only the effect of item was significant, $F(1,26) = 47.53$, $p < .01$, $R^2 = 0.151$, with better performance on the T/D comparisons (0.96) than on the T/S comparisons (0.87). No other effects or interactions were significant. However, we should mention that the Position × Delay × Item interaction approached significance, $F(1,26) = 2.84$, $p < .11$, $R^2 = 0.005$, with second-phrase T/S discrimination the only condition that did not remain roughly equal or decline over time, improving slightly from 0.84 to 0.87.

The median ratings of each stimulus type at each delay were subjected to a 2 Experience Levels × 2 Target Positions × 2 Time Delays × 3 Item Types ANOVA. The Delay × Item interaction was not significant; the data are shown in Table 2 for comparison to the other experiments. Again, significant effects that do not involve item type merely reflect some mixture of shifts in performance and in response criteria. The effect of target position was significant, $F(1,26) = 2.98$, $p < .05$, $R^2 = 0.003$. The effect of delay was significant, $F(1,26) = 11.11$, $p < .01$, $R^2 = 0.010$. The effect of item type

was significant, $F(2,52) = 793.01$, $p < .01$, $R^2 = 0.711$, indicating that listeners discriminated among the items (means of 4.8, 2.0, and 1.2 for median ratings of T, S, and D items respectively). No other effects or interactions were significant.

DISCUSSION

Overall, performance was much better here than in Experiment 1 (0.92 vs. 0.75), but there was no significant change in performance with the delay. We do not believe there is a ceiling effect here, at least not with the crucial T/S comparisons, because T/S performance was well below T/D (0.86 vs. 0.96). However, the crucial aspect concerned the improvement in memory over time. The removal of the intervening material in effect removed the conditions that facilitate improvement in T/S performance across the delay. Thus whatever might be the interfering role of the intervening material, its presence seems necessary to the improvement effect observed in Experiments 1 and 2. In addition, as will be seen in Experiment 4, filling the blank interval with meaningless interference, which often has a deleterious in verbal memory tasks (Baddeley, 1997), had virtually no effect in comparison with Experiment 3. This outcome supports the recoding hypothesis and leads us to reject the retrieval-strategy-shift hypothesis.

The most obvious difference in results between Experiment 3 and Experiment 1 is the overall difference in proportion correct. A plausible explanation attributes that difference to the removal of interference during the delay. However, we do not think that that is the best explanation. First, Experiment 3 removed not only the music during the delay interval but also the music preceding second-phrase targets. We think that this increased definition of the target was responsible for the generally higher performance, for several reasons. First, Dowling et al. (1995) simply omitted music during the delay (with already well-defined targets) in their Experiment 4 as opposed to Experiments 1 and 2, and they found performance unchanged with the empty delay. Second, we performed a pilot study like Experiment 3 but with the music preceding second-phrase targets retained, and we found overall performance about the same as in Experiments 1 and 2 (and much lower than in Experiment 3). That is, when listeners were presented with two potential targets on each trial, performance declined compared with a single unambiguous target presentation.

The musical structure and meaning account had predicted both worse performance overall as well as the disappearance of T/S improvement when the continuous music was removed. The results did not show the overall decrease in performance, but T/S improvement did disappear. Therefore we cannot rule out the possibility that structure and meaning have something to do with the facilitation of recognition.

Since the improvement over time in T/S performance observed in Experiments 1 and 2 disappeared when the continuous music was removed from the delay interval, we wished to find out whether this improvement was due to the complex character of the intervening material, or whether less complex musical patterns would be sufficient to facilitate improvement. Therefore in Experiment 4 we inserted a rhythmical but musically close-to-meaningless "oom-pah-pah" pattern into the delay.

Experiment 4

METHOD

Thirty listeners served in Experiment 4, of whom 16 were moderately experienced and 14 were inexperienced. Experiment 4 was the same as Experiment 3, except that the time interval between the presentation of the target and the test phrase was filled with an "oom-pah-pah" pattern on the beat, with a synthesized bass drum sound on the first beat and a synthesized woodblock sound on the second and third beats of every measure. This sounded more like "thump-click-click" than like "oom-pah-pah."

RESULTS

Area scores were subjected to a 2 Experience Levels × 2 Target Positions × 2 Time Delays × 2 Item ANOVA. The Delay × Item interaction was not significant; Table 1 shows the pattern of area scores for comparison to the other experiments. The effect of experience was significant, $F(1,28) = 4.78$, $p < .05$, $R^2 = 0.054$, with the more experienced listeners performing better (0.92 vs. 0.85). The effect of item was significant, $F(1,28) = 56.82$, $p < .01$, $R^2 = 0.102$, with better performance on the T/D comparisons (0.94) than T/S (0.84). And the interaction of Experience × Item Type was significant, $F(1,28) = 8.10$, $p < .01$, $R^2 = 0.015$, with T/S performance distinctly worse than T/D performance for inexperienced listeners (0.79 vs. 0.92), but not as much worse for experienced listeners (0.89 vs. 0.96). No other effects or interactions were significant. However, the Position × Delay interaction approached significance, $F(1,28) = 3.62$, $p < .07$, $R^2 = 0.013$, with second-phrase performance improving slightly over time (0.87 to 0.90) although first-phrase performance declined (0.92 to 0.87).

The median ratings of each stimulus type at each delay were subjected to a 2 Experience Levels × 2 Target Positions × 2 Time Delays × 3 Item Types ANOVA. The Delay × Item interaction was significant, $F(2,56) = 3.30$, $p < .05$, $R^2 = 0.004$, with hits and S false-alarm ratings declining over time and D false alarms remaining steady (see Table 2). Note that this pattern corresponds to both T/S and T/D discrimination becoming somewhat worse (see Table 1). Again, significant effects that do not involve item type merely reflect some mixture of shifts in performance and in response criteria.

The effect of target position was significant, $F(1,28) = 4.36$, $p < .05$, $R^2 = 0.002$, as was the effect of delay, $F(1,28) = 6.86$, $p < .02$, $R^2 = 0.006$. The effect of item type was significant, $F(2,56) = 275.05$, $p < .01$, $R^2 = 0.687$, indicating that listeners discriminated among the items (means of 4.9, 2.2, and 1.2 for median ratings of T, S, and D items, respectively). The interaction of Experience × Item was significant, $F(2,56) = 4.06$, $p < .05$, $R^2 = 0.010$, reflecting the better performance by more experienced listeners observed with the area scores. The Position × Item interaction was significant, $F(2,56) = 4.15$, $p < .05$, $R^2 = 0.004$, with first-phrase S lures easier to reject than second-phrase S lures (ratings of 1.9 vs. 2.4). No other effects or interactions were significant.

DISCUSSION

As in Experiment 3, there were no significant changes in performance over time. In fact, as can be seen in Tables 1 and 2, the overall results for the two experiments were virtually identical. This suggests that, whatever is facilitating the improvement seen in Experiments 1 and 2, it probably depends on the presence of relatively complex musical patterns during the delay. This suggestion is compatible with both the recoding and musical structure hypotheses. The oom-pah-pah pattern of Experiment 4 was sufficiently different from the target so that it did not evoke it and bring about recoding, and it did not serve as a meaningful continuation of the music. These results further persuade us to reject any simple interference hypothesis. Silence presumably causes less interference than the metrical filler, but the results for the empty and metrical-filler delays were virtually identical.

This is the only one of the five experiments in which we found effects of experience. Our moderately experienced listeners were not as distracted by the oom-pah-pahs as our inexperienced listeners, for whom it had a greater impact on T/S discrimination than on T/D. In Experiment 1, inexperienced and moderately experienced listeners were alike in their response to the more ecologically valid continuous music.

Now we can consider the possibility that the prospect of being tested at the end of the trial is an important ingredient of the improvement in recognition observed in Experiments 1 and 2, as Erdelyi (1996) had suggested. In Experiment 5, we led listeners to believe that they were simply rating the pieces for attractiveness. Then at the end of the first trial we asked them to perform a recognition test. We constructed the trials using the same materials as Experiment 2.

Experiment 5

Since each listener could participate in only one trial, we had to be selective concerning the conditions we could run. Given the results of Experi-

ments 1 and 2, we decided in Experiment 5 to focus on T/S performance with second-phrase targets. In the interests of efficiency, we did not include conditions involving first-phrase targets or D test items that were not as pertinent to the improvement in which we were interested.

METHOD

Sixty listeners served in Experiment 5, of whom 34 were moderately experienced and 26 inexperienced. Experiment 5 used T and S stimuli from List 1 of Experiment 2, but each listener heard just one stimulus. Listeners participating in individual sessions (in connection with another unrelated study) were led to believe that they were about to rate a series of stimuli for "liking" or "pleasantness." However, after the first stimulus, the experimenter interrupted the task and asked the listener to judge whether the last phrase heard had occurred earlier in the piece, using the same 6-point response scale as in the other experiments. Two second-phrase T and two second-phrase S items from each of the three delay conditions in List 1 of Experiment 2 were selected for Experiment 5 on the basis of median ratings. We used those stimuli closest to the overall median rating in each condition. Listeners were randomly assigned to condition, and stimuli were randomly assigned with the constraint that each stimulus appear equally often.

RESULTS

Listeners' ratings were subjected to a 3 Delay × 2 Items ANOVA. There were no significant results, though the effect of item type approached significance, $F(1,54) = 2.75$, $p < .11$. Table 3 shows the mean ratings of the stimuli and includes mean ratings of the same stimuli from Experiment 2 for comparison.

DISCUSSION

As can be seen in Table 3, T/S discrimination in Experiment 5 was best at the 4-s delay and fell off after that to a level not appreciably better than chance. This is contrary to the improvement over time seen in Experiment

TABLE 3
Means of Listeners' Median Ratings of Second-Phrase T and S Stimuli at Three Time Delays in Experiment 5, Compared with Ratings of Those Stimulus Types from Experiment 2

	Delay (s)		
	4	15	30
Experiment 5			
T	4.4	4.0	4.0
S	3.1	3.8	3.8
Experiment 2			
T	4.4	4.2	4.0
S	3.2	2.8	2.6

2. We do not believe that higher within-cell variability led to the negative results of Experiment 5: standard deviations for the six cells in Experiment 5 ranged from 0.84 to 1.49; and for the corresponding six cells in Experiment 2 for second-phrase items, ranged from 1.02 to 1.46. We are inclined to conclude from these results that the prospect of testing is a necessary ingredient to the recognition improvement observed in Experiments 1 and 2. However, caution may be advisable. The shift in task communicated to listeners at the end of the single trial was itself disruptive. That it was not totally disruptive of memory can be seen in the good performance at 4 s. However, it is clear that a memory task performed in retrospect after a somewhat jarring violation of task expectancy is not the same as a memory task performed smoothly in the context of a sequence of similar trials. Therefore we can view the present result as suggesting the expectation of a test as a factor in producing recognition improvement, but not as conclusive.

General Discussion

This series of experiments demonstrates an improvement in memory over time in T/S discrimination with music in circumstances that approximate a natural listening situation more closely than in previous studies. This improvement appeared stronger for second-phrase targets than for opening phrases in the minuets. The improvement occurred when complex intervening stimuli—the natural continuation of the piece—were present between target and test (Experiments 1 and 2). Although those competing musical patterns may have interfered somewhat with recognition performance, nevertheless their presence was necessary to the improvement effect, in which listeners gain in their ability to reject lures resembling targets in broad outline but differing in detail (see Figure 2). Perfunctory, contentless patterns, such as the oom-pah-pah pattern of Experiment 4, do not facilitate improvement any more than a simply blank interval. That is, the important thing is not that the delay be filled, but that it be filled with musically meaningful material that engages the listener.

IMPLICIT VS. EXPLICIT, AUTOMATIC VS. CONTROLLED

These results showing retention of detail contrast with those of Sachs (1967), for example, who found that, after a delay filled by reading the continuation of a paragraph, adults lost track of surface details of a sentence but retained the gist and its meaning. (See Gernsbacher, 1985, for an extensive review of similar findings, and Goldinger, 1996; Goldinger, Kleider, & Shelley, 1999, for reviews of results in which surface details are retained.)

When we embarked on this series of studies, we fully expected results converging with those of Sachs, whose experimental paradigm was closely parallel to ours. Clearly our results diverge. However a comment of Goldinger's (1996, p. 1166) may help to solve the puzzle. He notes not only that "memory for surface details is more often revealed by implicit measures than by explicit measures," but that "the passage of time differentially affects each measure: Surface details rapidly fade from explicit memory, but persist in implicit memory." Explicit memory involves the sort of case in which you can literally bring to mind the target in question—essentially to recall it (Tulving, 1985). Implicit memory involves the assessment of performance on tasks in which memory for the target in question, although not being explicitly recalled, nevertheless affects behavior in some measurable way. Explicit memory is a "cognitively controlled" process—we can recall the memory at will or dismiss it from consciousness. Implicit memory will appear whether we want it to or not—it is automatic. Dowling et al. (1995) argued that T/S discrimination was an automatic process. Harking back to the inability of Dowling and Barlett's (1981) listeners to respond "old" to S lures, they say:

> We believe the evidence suggests that the encoding of contours is largely a controlled process, and the encoding of pitch interval pattern information is largely automatic. The critical evidence for this involves interference. . . . T/D discrimination is subject to interference from concurrent tasks, and T/S discrimination is not. (Dowling et al., 1995, p. 147)

We argue in the discussion of Experiment 2, earlier, that the processing of second-phrase targets, in comparison with that of first-phrase targets, is more likely to be automatic. The automaticity of T/S discrimination converges with the suggestion that it is indicative of implicit memory. If T/S discrimination is an automatic process, it is perhaps not so surprising that performance improves over time.

Further suggestive support for the notion that the memory processes we are observing are automatic comes from the study of Wright et al. (1985) discussed earlier. They obtained the same results with pigeons, monkeys, and humans. The pigeons in particular must be using automatic, implicit processes of retrieval, as distinct from explicit, controlled processes.

One way of putting this suggestion is to say that the recognition improvement we have observed occurs because, while listening continues, the processing of already-heard material proceeds automatically. That processing has the effect of increasing the precision of the memory representation of what was heard earlier. We should not expect, however, that that processing will occur purely implicitly and automatically. Kinoshita (2001), for example, has pointed out how explicit processes can impinge even on tasks

in which the instructions call for implicit processing. In the present tasks, the instructions suggest explicit processing, and so the actual result for our listeners must be some mixture of the two.

One aspect that these studies leaves unresolved is the importance of meaning and musical structure of the intervening material. Although it is true that improvement disappeared when we included empty and meaningless material during the delay, we are left asking what the role of that material is; for example, is it to engage the listener's attention, or to aid the listener's understanding of the connection of the target and the test? This requires further study.

LISTENING TO MUSIC

The present experiments allow us to study the early stages of formation of memories for exact details of phrases of music. Taking these experiments together with previous results (Dowling, 1978; Dowling & Bartlett, 1981; DeWitt & Crowder, 1986; Dowling et al., 1995, 1999), we can conclude that the listener does not initially remember exactly what was heard, but remembers certain global features of the overall pattern, such as contour and key. However, with additional automatic processing during the presentation of additional musical material, the memory trace becomes more and more precise over a period of up to 4 min. These improvements are similar to those reviewed by Erdelyi (1996). But a clear implication of the improvements in T/S discrimination in recognition memory is that there are qualitative changes over time in the contents of memory.

The novelist Marcel Proust is often cited in connection with the importance of implicit, procedural memory in our life experience (Baddeley, 1997) and is sometimes held up as a proponent of veridical and unchanging memory retrieval (Eakin, 2000). In fact, during the 1890s, Proust was an avid student of the philosopher Henri Bergson and his groundbreaking approaches to aspects of memory (Painter, 1959). So it is perhaps not surprising that Proust has provided us with a vivid description of the initial stages of encoding and recoding of a memory of a novel piece of music. In *Swann's Way* (*Du côté de chez Swann*), Swann has just been struck by a particular phrase in the Andante of Vinteuil's (Fauré's) First Violin Sonata:

> The notes we hear . . . with their pitches and durations . . . cover surfaces of varied dimensions before our eyes, tracing arabesques and giving us sensations of size, of continuity, of stability, of caprice. But the notes have vanished before these sensations are well enough formed in us to avoid being submerged by following (or even simultaneous) notes. And that impression continues to envelop in its liquid background motifs that at moments emerge from it, hardly distinguishable, only to dive back and disappear at once, motifs known only by the particular plea-

> sure they provide—impossible to describe, to recall, to name—ineffable. It was as if memory, like a worker striving to erect a solid foundation in the midst of a flood, while making us facsimiles of these fleeting phrases, would not allow us to compare them to those that follow, and to differentiate them. Thus hardly had the delicious sensation Swann felt expired, but his memory gave him a provisional and summary transcript of it even while he continued to listen. He took a good enough look at the transcript while the piece continued, so that when the same impression suddenly returned, it was no longer impossible to grasp. (Proust, 1999, p. 173, author's translation)

This account touches on two aspects of early memory processing relevant to the present discussion. First, it suggests that memory processing of previously presented information continues even while new information is entering the system. This is what we suggested above in terms of continued automatic processing. Second, the listener's experience of the piece is in a continual state of flux. If the listener reflects on what was heard, a different representation is retrieved depending on the time elapsed since it was heard initially. Proust suggests that the initial experience is of global, diffuse properties of a phrase that are difficult to encode. This is followed by the abstraction of general features—a "provisional and summary transcript." Proust suggests that that abstraction aids later recognition when the listener is again cued with the overall impression. However, experiments convince us that the story does not stop there. The processing of a phrase, once heard, continues automatically even while the listener hears new phrases. What the listener can remember having heard continually changes during continued listening.

We shall close with some considerations about musical form, time, and the listener's experience. Form provides a structure for time; it divides time into segments, providing a pattern that can be mapped in the listener's mind. In giving structure to time for the listener, musical form serves to direct expectancies, and hence the trajectory of processing of incoming information. As Jones (1981) points out, listeners

> generate subjective space-time paths . . . in response to certain features of the external stimulus pattern. These mental "paths" function as psychological expectancies. And it is through extrapolation of these mental spatio-temporal patterns that a person comes to anticipate "where" in space [pitch] and "when" in time future events may occur. Expectancies, at least initially, are typically ideal or simplified paths. They are continuous, rhythmically generated paths that allow us to guide our attention to approximately correct neighborhoods. But what is most important is that organisms possess subjective generators that resemble those outlined in the representation of world patterns. (Jones, 1981, p. 571)

It is very likely that not only the content of the intervening material between a target and its test is important (compare Experiments 1 and 4), but also its musical structure. In the present studies, musically untrained listeners showed the same implicit sensitivity to aspects of musical structure displayed by moderately trained musicians. (Note that the performance of the two groups only diverged in Experiment 4, which interpolated meaningless, unstructured material.) We thus believe that the structure of our materials is accessible to the musically untrained. A group of additional studies will explore the importance of structure by systematically manipulating structural aspects of the intervening material in order to isolate the features that are important for the memory improvement effect (Dowling & Tillmann, in preparation).

The qualitative changes in memory representation have implications for musical experience. While we should be cautious concerning the extrapolation of these results to situations in which listening is not followed by a test, it may be that simply becoming attentively absorbed in the music, like Proust's Swann, may be sufficient to trigger the pattern observed here. Such qualitative changes will have an impact on the ways in which new material in a piece is experienced in relation to old material, leading to differences in the experience of similarity and difference as the piece progresses. And they will also affect how expectancies are generated, and the degree to which those expectancies are experienced as fulfilled or violated.

Music provides us with a domain in which we can study listeners' representations of the structuring of time, both because that structure is well-defined in music and because musical patterns succeed in holding the listener's interest and engaging the automatic brain processes in perception and memory whose secrets we seek to disclose.[1]

References

Baddeley, A. (1997). *Human memory: Theory and practice*. Boston: Allyn & Bacon.

Beethoven, L. van (1967). *Klaviersonaten,* vol. 1. Munich: G. Henle.

Beethoven, L. van (1987). *Bagatelles, rondos, and other shorter works for piano.* New York: Dover.

Beethoven, L. van (1990). *Tänze für Klavier.* Munich: G. Henle.

DeWitt, L. A., & Crowder, R. G. (1986). Recognition of novel melodies after brief delays. *Music Perception, 3,* 259–274.

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

1. We thank Howard Medlock for assisting in stimulus preparation and collecting data, Hervé Abdi, James C. Bartlett, and Mari Riess Jones for helpful comments, and Emmanuel Bigand for contributions to the initial impetus for testing memory by means of continuous passages of classical music.

Dowling, W. J., Barbey, A., & Adams, L. (1999). Melodic and rhythmic contour in perception and memory. In S. W. Yi (Ed.), *Music, mind, and science* (pp. 166–188). Seoul: Seoul National University Press.
Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology, 1,* 30–49.
Dowling, W. J., & Harwood, D. L. (1986). *Music cognition.* Orlando, FL: Academic Press.
Dowling, W. J., Kwak, S.-Y., & Andrews, M. W. (1995). The time course of recognition of novel melodies. *Perception & Psychophysics, 57,* 197–210.
Dowling, W. J., & Tillmann, B. (in preparation). *Musical structure and memory.*
Eakin, P. J. (2000). Autobiography, identity, and the fictions of memory. In Schacter, D., & Scarry, E. (Eds.) *Memory, belief, and brain* (pp. 290–306). Cambridge, MA: Harvard University Press.
Erdelyi, M. H. (1996). *The recovery of unconscious memories: Hypermnesia and reminiscence.* Chicago: University of Chicago Press.
Estes, W. K. (1999). Models of human memory: A 30-year retrospective. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections* (pp. 59-86). Mahway, NJ: Lawrence Erlbaum.
Frey, P. W., & Adesman, P. (1976). Recall memory for visually presented chess positions. *Memory & Cognition, 4,* 541–547.
Gabrielsson, A. (1999). The performance of music. In D. Deutsch (Ed.), *The psychology of music* (pp. 501–602). San Diego, CA: Academic Press.
Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology, 17,* 324–363.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model of both recognition and recall. *Psychological Review, 91,* 1–67.
Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 1166–1183.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 104,* 251–279.
Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition, 27,* 328–338.
Haydn, J. (1984). *Complete piano sonatas,* 2 vols. New York: Dover.
Haydn, J. (1989). *Tänze für Klavier,* Hob. IX: No. 3, 8, 11, 12. Vienna: Schott.
Jones, M. R. (1981). Only time can tell: On the topology of mental space and time. *Critical Inquiry, 7,* 557–576.
Kinoshita, S. (2001). The role of involuntary aware memory in the implicit stem and fragment completion tasks: A selective review. *Psychonomic Bulletin & Review, 8,* 58–69.
Lakos, A. (Ed.) (1994). *Alte Tänze.* Budapest: Kunemann.
Mozart, W. A. (1956). *Sonatas and fantasies for the piano.* Bryn Mawr, PA: Theodore Presser.
Mozart, W. A. (1992). *Sonatas and variations for violin and piano,* 2 vols. New York: Dover.
Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104,* 839–862.
Murdock, B. B. (1999). The buffer 30 years later: Working memory in a theory of distributed associative memory (TODAM). In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections* (pp. 35–57). Mahway, NJ: Lawrence Erlbaum.
Painter, G. D. (1959). *Marcel Proust: A biography,* 2 vols. London: Chatto & Windus.
Proust, M. (1999). *A la recherche du temps perdu* (edition in 1 vol.). Paris: Gallimard.
Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2,* 437–442.
Schubert, F. (1989). *Dances for solo piano.* New York: Dover.

Shepard, R. N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology, 62,* 302–309.
Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182,* 990–1000.
Tillmann, B., & Bigand, E. (1998). Influence of global structure on musical target detection and recognition. *International Journal of Psychology, 33,* 107–122.
Tulving, E. (1983). *Elements of episodic memory.* Oxford: Oxford University Press.
Tulving, E. (1984/1986). Precis of *Elements of episodic memory* with commentary. *Behavioral & Brain Sciences, 7,* 223–268; *9,* 566–577.
Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1–12.
Tulving, E. (1987). Multiple memory systems and consciousness. *Human Neurobiology, 6,* 67–80.
Tulving, E., & Watkins, M. J. (1975). Structure of memory traces. *Psychological Review, 82,* 261–275.
Wright, A. A., Santiago, H. C., Sands, S. F., Kendrick, D. F., & Cook, R. G. (1985). Memory processing of serial lists by pigeons, monkeys, and people. *Science, 229,* 287–289.