Research

They saw a movie: Long-term memory for an extended audiovisual narrative

Orit Furman,[1] Nimrod Dorfman,[1] Uri Hasson,[2] Lila Davachi,[2,3] and Yadin Dudai[1,4]

[1]*Department of Neurobiology, The Weizmann Institute of Science, Rehovot 76100, Israel;* [2]*Center for Neural Science, New York University, New York, New York 10003, USA;* [3]*Department of Psychology, New York University, New York, New York 10003, USA*

We measured long-term memory for a narrative film. During the study session, participants watched a 27-min movie episode, without instructions to remember it. During the test session, administered at a delay ranging from 3 h to 9 mo after the study session, long-term memory for the movie was probed using a computerized questionnaire that assessed cued recall, recognition, and metamemory of movie events sampled ~20 sec apart. The performance of each group of participants was measured at a single time point only. The participants remembered many events in the movie even months after watching it. Analysis of performance, using multiple measures, indicates differences between recent (weeks) and remote (months) memory. While high-confidence recognition performance was a reliable index of memory throughout the measured time span, cued recall accuracy was higher for relatively recent information. Analysis of different content elements in the movie revealed differential memory performance profiles according to time since encoding. We also used the data to propose lower limits on the capacity of long-term memory. This experimental paradigm is useful not only for the analysis of behavioral performance that results from encoding episodes in a continuous real-life-like situation, but is also suitable for studying brain substrates and processes of real-life memory using functional brain imaging.

Experimental protocols that probe brain correlates of episodic memory formation commonly use paradigms in which memoranda are presented as individual items devoid of continuous context outside of the laboratory setting (Winocur and Weiskrantz 1976; Buckner et al. 2000). In contrast, real-life episodic memory is the result of ongoing encoding within a highly contextualized and dynamically changing perceptual, cognitive, and affective framework (Tulving 1983, 2002; Suddendorf and Busby 2005). Though the importance of real-life conditions in memory research has long been recognized (Neisser 1978; Cohen 1996), it is rather difficult to harness its naturalistic attributes in controlled, reproducible laboratory settings (Dudai 2002). Using movies as stimulus material can remedy some of these difficulties.

Movies are capable of simulating aspects of real-life experiences by fusing multimodal perception with emotional and cognitive overtones (Eisenstein 1969; Morin 2005). They also permit controlled, reproducible presentation of continuous, contextualized, and dynamic sets of stimuli to-be-remembered, and selection of cognitive and affective types of content. The use of cinematic material to probe memory can be traced to the early days of cinema (Boring 1916), but did not catch on, a few exceptions notwithstanding (Beckner et al. 2006). Realizing the potential advantage of movies as multimodal stimuli on-the-go, Hasson et al. (2004) used a trade fiction movie to analyze brain circuits that process perceptual and affective information while attending the ongoing cinematic narrative, and unveiled correlated spatiotemporal brain activation patterns in multiple subjects while watching identical scenes.

Here, we describe the use of a 27-min narrative movie to investigate long-term cued recall and recognition as well as metamemory judgments. We measured memory performance of several groups of participants, each at a different delay ranging from 3 h to 9 mo after watching the movie, by probing memory for events occurring in the movie ~20 sec apart. Participants remembered many events in the movie that they had seen only once without prior instructions to remember it, even months after watching it. We have dissected multiple facets of memory and metamemory as a function of time and type of occurrence in the movie. Our analysis also suggests lower limits on the capacity of long-term memory for a real-life-like situation.

Results

Memory of the movie persists for months

The first set of experimental groups watched the movie and answered the questionnaire (Fig. 1; see Materials and Methods) once, each at a selected time interval, up to 9 mo after the study session. Memory performance (quantified as correct answers across all confidence levels) declined as a function of the time between study and test (Fig. 2A) ($F_{(5,58.23)} = 26.07$, $P < 0.0001$). The No-Movie group performed at chance level on a two-alternative forced choice test (53% ± 2% correct). Using the No-Movie group as an estimate of chance memory performance, all groups except the 9-mo group performed above chance (No-Movie vs. 3 h $T_{(61.32)} = 8.21$ $P < 0.0001$; No-Movie vs. 1 wk $T_{(63.04)} = 8.45$, $P < 0.0001$; No-Movie vs. 3 wk $T_{(50.63)} = 7.04$, $P < 0.0001$; No-Movie vs. 3 mo $T_{(45.44)} = 4.6$, $P < 0.0005$; No-Movie vs. 9 mo $T_{(44.37)} = 2.37$, $P < 0.3$). Although the 9-mo group showed only a trend for difference from the No-Movie group in this analysis,[5] it did show distinct differences in performance from the No-Movie group on additional metamemory measures (see below; Fig. 3A).

Heuristic subdivisions in long-term memory

Pairwise comparisons among all groups, corrected for multiple comparisons, reveal superior performance of the 3-h, 1-wk, and 3-wk groups compared with the 3-mo and 9-mo groups (Fig. 2A) (3 h vs. 1 wk, $T_{(84.19)} = -0.28$, $P = 1$; 3 h vs. 3 wk, $T_{(74.02)} = 2.18$,

[4]**Corresponding author.**
E-mail yadin.dudai@weizmann.ac.il; fax (972) 8-946-9244.
Article is online at http://www.learnmem.org/cgi/doi/10.1101/lm.550407.

[5]When compared in isolation using a two-tailed *t*-test, the difference between the 9-mo and the No-Movie group was significant, $P = 0.017$.

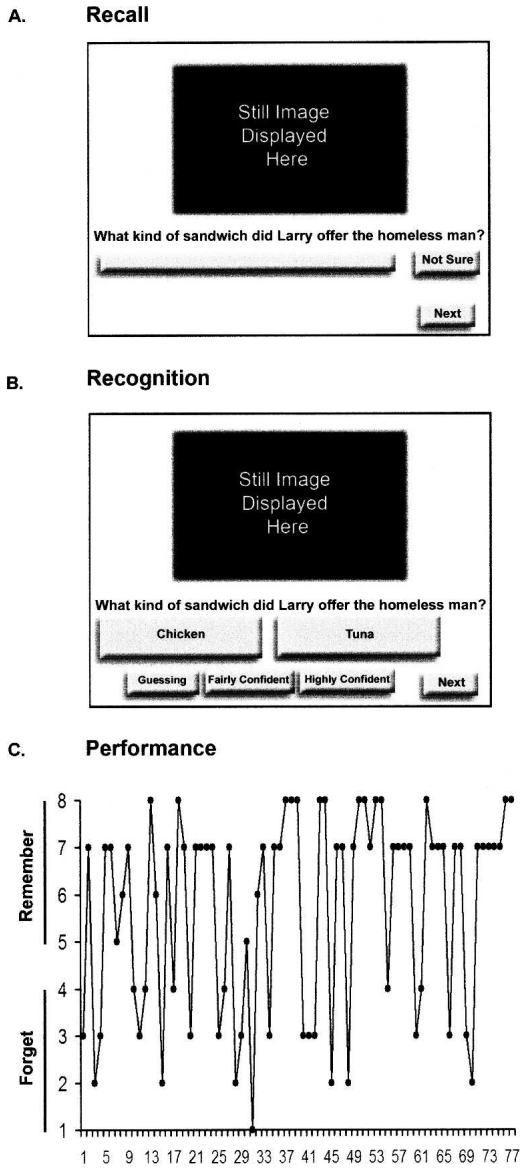

Figure 1. Memory performance was measured by an interactive computerized questionnaire. (*A*) Recall. Each question was accompanied by a relevant snapshot from the movie, intended to serve as a visual cue. If the answer is known, the participant types it and proceeds to the next question by pressing "Next." Otherwise, pressing the button marked "Not Sure" leads to the recognition phase. (*B*) Recognition. The same question and visual cue are shown on the screen. The participant is instructed to choose between two alternatives answers, and to rate her/his level of confidence as either "highly confident," "fairly confident," or "guessing." Both elements are required before proceeding to the next question. (*C*) Performance. Results from a single participant, tested 3 wk after watching the movie, are presented as an example. Response for each question is coded according to two parameters: correct/incorrect and level of confidence (recall, high-confidence recognition, low-confidence recognition, guessing). The arbitrary rating units indicate an integration of these parameters as follows: (8) correct recall, (7) correct high-confidence recognition, (6) correct low-confidence recognition, and (5) correct guessing. Levels indicating forgotten facts are: (4) incorrect recall, (3) incorrect high-confidence recognition, (2) incorrect low-confidence recognition, and (1) incorrect guessing.

$P < 0.4$; 3 h vs. 3 mo, $T_{(71.07)} = 5.12$, $P < 0.0001$; 3 h vs. 9 mo, $T_{(66.46)} = 6.67$, $P < 0.0001$; 1 wk vs. 3 wk, $T_{(76.31)} = 2.48$, $P < 0.3$; 1 wk vs. 3 mo, $T_{(73.61)} = 5.41$, $P < 0.0001$; 1 wk vs. 9 mo, $T_{(68.63)} = 6.94$, $P < 0.0001$; 3 wk vs. 3 mo, $T_{(57.18)} = 3.34$, $P < 0.05$; 3 wk vs. 9 mo, $T_{(53.89)} = 5.23$, $P < 0.0001$; 3 mo vs. 9 mo $T_{(47.56)} = 2.35$, $P < 0.3$). Because additional performance measures detailed below indicated significant differences between but not within the hours-to-weeks groups vs. the months groups, we introduced for convenience the notations Shorter-Time-Interval (ST) and Longer-Time-Interval (LT) to refer to these groups, accordingly. This division, however, does not necessarily imply a categorical distinction between ST and LT groups, as they could still be placed on a continuum (see Discussion).

Recall attempts over time

Participants made fewer attempts at recall as more time passed between study and test sessions (Fig. 2B) ($F_{(4,52.27)} = 3.09$, $P < 0.025$). Pairwise contrasts did not reveal significant differences between the various ST groups in attempts at using recall.

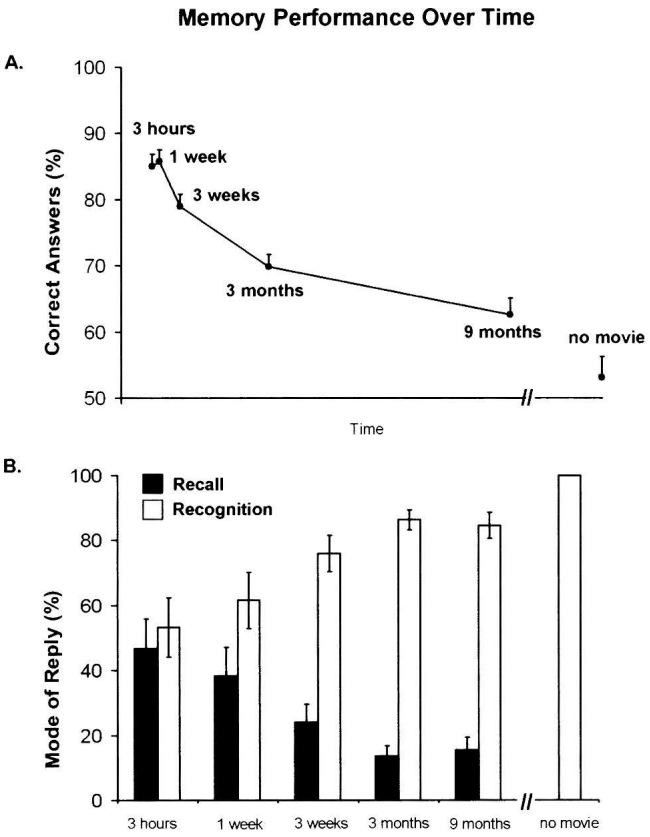

Figure 2. Memory performance over time. (*A*) Forgetting curve. Each group was tested only once at the indicated time points (3 h, $n = 8$; 1 wk, $n = 8$; 3 wk, $n = 12$; 3 mo, $n = 17$; 9 mo, $n = 12$). Data are presented as percent of correct answers overall (collapsing across difference confidence levels). A significant statistical difference is found between performance of Shorter-time (ST) groups (3 h, 1 wk, and 3 wk) and Longer-time (LT) groups (3 and 9 mo); for the rationale for this notational classification, see Results. Performance declined significantly between, but not within ST and LT groups. In the No-Movie protocol (control group in which the subjects did not see the movie, $n = 8$), correct answers overall were around 50%, i.e., chance performance. (For statistics, see Results.) (*B*) Mode of reply, presented as percent of recall (black) or recognition (white) answers of total answers (correct + incorrect). A significant decrease in percent recall is accompanied by an increase in percent recognition over time. Significant differences in percent of recall are detected between: 3 h and 3 mo, 3 h and 9 mo, 1 wk and 3 mo. No significant difference is detected between 3 and 9 mo. In the No-Movie protocol, all questions were answered using recognition. Values are mean ± SEM.

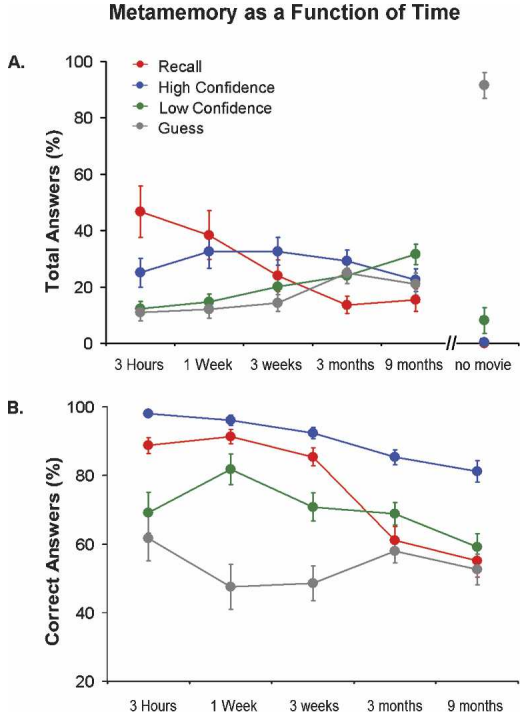

Figure 3. Metamemory as a function of time. Answers were coded according to four levels of confidence: recall (considered as the highest confidence level, red), high-confidence recognition (HCR, blue), low-confidence recognition (LCR, green), and guessing (gray). (*A*) Proportion of answers (correct + incorrect) as a function of confidence levels. Each curve depicts the mean percent of answers given using each confidence level, per time-interval group. Significant decline is seen for recall, no difference is detected for HCR answers, and a significant increase is found for LCR and guess answers. In the "No-Movie" protocol, 92% of the answers were guesses. (For statistics, see Results.) (*B*) Proportion of correct answers as function of confidence levels across time groups. Each curve depicts the mean proportion of correct answers of answers given in each confidence level, per time interval. Significant decreases are found in all curves except correct guesses. Values are mean ± SEM.

However, recall attempts were higher in the 3-h group compared with the 3- and 9-mo groups ($P < 0.005$ and $P < 0.025$, respectively), and were higher in the 1-wk group as compared with the 3-mo group ($P < 0.03$).

Memory confidence over time

The proportions of all answers made at different levels of confidence were compared across time-interval groups, revealing a surprisingly stable proportion of high-confidence recognition (HCR) answers over time, and a significant decrease in the use of recall over time. In contrast, the proportion of low-confidence recognition (LCR) and guess answers significantly increased over time (Fig. 3A; see also Fig. 4B for comparison between numbers of correct and incorrect answers) (Recall: $F_{(4,38.97)} = 5.35$, $P < 0.0002$; HCR: $F_{(4,47.37)} = 0.9$, $P < 0.47$; LCR: $F_{(4,53.23)} = 6.03$, $P < 0.0005$; Guess: $F_{(4,52.27)} = 3.09$, $P < 0.03$).

One might expect a smaller proportion of correct answers as confidence declines with the passage of time. To test this assumption, the proportion of correct answers was calculated from all answers made in each level of confidence (Fig. 3B). A two-way analysis of the proportion of correct answers unveiled significant effects for level of confidence ($F_{(3,4369)} = 77.16$, $P < 0.0001$), time-interval group ($F_{(4,51.83)} = 11.65$, $P < 0.0001$), but also for their interaction ($F_{(12,4369)} = 7.48$, $P < 0.0001$). The main effect of confidence level is that, indeed, a significant decline over time in correct answers is seen for all confidence levels, except guessing (correct recall: $F_{(4,203.6)} = 21.45$, $P < 0.0001$; correct HCR: $F_{(4,421.5)} = 8.42$, $P < 0.0001$; correct LCR: $F_{(4,195.9)} = 3.05$, $P < 0.02$; correct guess answers: $F_{(4,183.5)} = 1.18$, $P < 0.3$).

Furthermore, the main group effect suggests that distribution of correct answers between the four possible levels of confidence is uneven among time-interval groups. The interaction between these two main effects, time and confidence-level, was further explored by performing pairwise contrasts, revealing performance differences that further support differentiation of ST and LT groups. As can be seen in Figure 3B, the ST groups have higher proportions of correct answers made using recall and HCR relative to the LT groups. This difference between time-interval groups is also detected in the analysis performed on correct responses by content clusters (see below).

The findings described above indicate that confidence measures of memory are time-sensitive: more high-confidence answers (using recall and HCR) are used after short time durations, while more lower-confidence answers (LCR and guessing) are used after longer time durations. We further sought to characterize the temporal dynamics of metamemory measures using analysis of answers by confidence level. We find that on the one hand, proportion of overall use of recall declines over time (Fig. 3A), as does the proportion of correct answers made using recall (Fig. 3B). On the other, overall proportion of HCR does not change significantly over time (Fig. 3A), and only a rather moderate decline is seen in the proportion of correct HCR answers over the entire 9-mo period (98% correct after 3 h, >80% correct after 9 mo). As for lower confidence levels (LCR and guess) (Fig. 3A), an increase is seen in overall proportion of answers made, while the proportion of correct answers decreases or remains unchanged over time (Fig. 3B). All in all, when subjects are allowed to freely choose between recall and recognition as a mode of reply, recall is used mainly while memory is recent, while HCR is used similarly for all retention durations, and use of LCR (as well as guesses) increases at longer intervals, compensating for the decline in recall use. The usefulness of probing metamemory is demonstrated as HCR answers are more often correct than less-confident recognition answers (namely, LCR and guesses) (Fig. 3B).

Memory density

The availability of memory performance scores that sample events every 20 sec in a 27-min episode renders it tempting to attempt tapping into long-term memory capacity per unit time, or density (Dudai 1997). Any such attempt is bound to yield only rough estimates. First, formal units that might be used for quantifying the stored memory, such as bits (Dudai 1997), are impossible to determine. Second, the questions that we devised only sample the occurrences in the study material. Third, the test protocol may not necessarily provide optimal retrievability for these sampled items. And fourth, the items retrieved are not necessarily independent, because the narrative of the movie might link them to each other. All of these confounds notwithstanding, because of the scarcity of estimates of long-term memory (Dudai 1997), we did consider this mental exercise informative and worthwhile.

Using the algorithm specified in the Materials and Methods, we estimate that 56% of information in the movie is remembered after 3 h, while 53%, 39%, 25%, and 19% are remembered after 1 wk, 3 wk, 3 mo, and 9 mo, respectively. Equating for the sake of calculation a memory unit as a questionnaire item to be answered correctly and assuming independence among items,

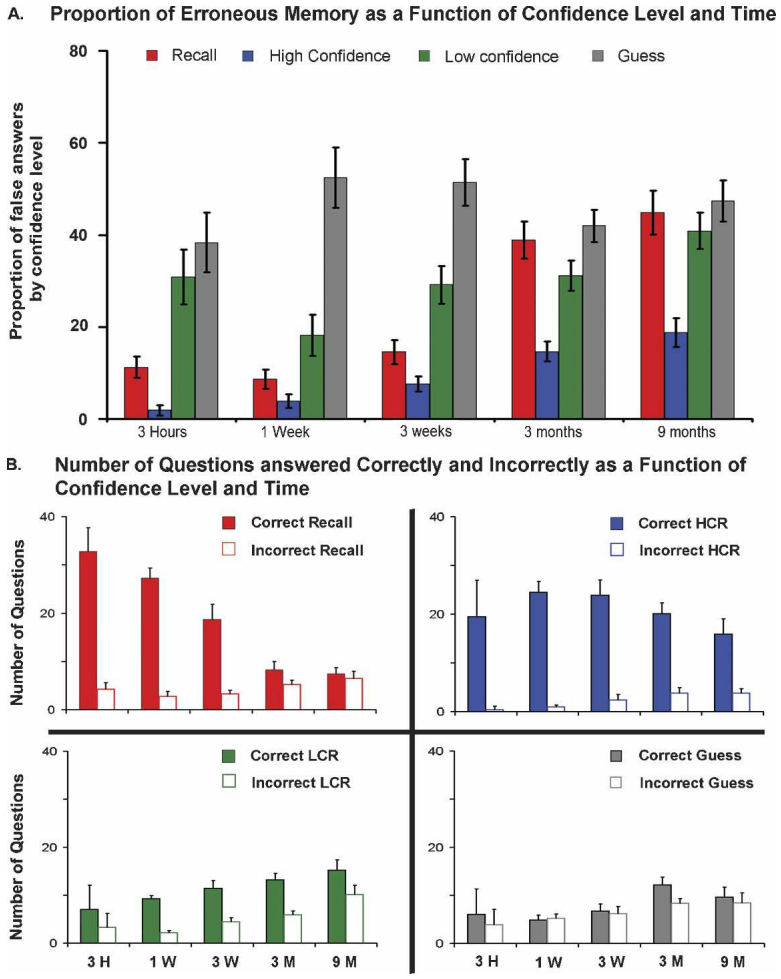

Figure 4. Erroneous memory as a function of confidence and time. Proportions of incorrect answers were calculated separately for each time interval group and confidence level (red, recall; blue, high-confidence recognition [HCR]; green, low-confidence recognition [LCR]; gray, guessing). (*A*) Proportions were calculated from total answers made with each confidence level, per time-interval group. A significant increase in proportion of incorrect answers is seen for all confidence levels. It is noteworthy that the least change is seen for HCR answers and not for recall answers, which were thought to represent the highest degree of confidence. Proportion of incorrect recall and HCR answers was significantly smaller for Shorter-time groups (3 h, 1 wk, 3 wk) than for Longer-time groups (3 and 9 mo). (*B*) Average number of correct and incorrect answers was calculated separately for each confidence level and time-interval group: recall (*top left*), HCR (*top right*), LCR (*bottom left*), and guess (*bottom right*). Values are mean ± SEM.

hence, 77 items that could be potentially encoded over 27 min, this implies retrievability ranging from 1.6 items per minute movie after 3 h to 1.4 and 1.0 items per 2 min after 3 and 9 mo, respectively. The robustness of the assumptions involved and the relevance to previous estimates of long-term memory capacity are discussed below.

Recall and recognition as function of memory content

Six independent raters were asked to classify the questions into eight predetermined categories (while allowing overlap of categories). Questions were then grouped into nonoverlapping clusters based on 66% agreement among raters. Clusters included plot themes, social interactions, couple relationship, and jokes and minor details (13, 19, 6, 5, and 14 questions, respectively). "Social interactions" is a nonoverlapping broader category than "couple relationship," and did not include questions relating to interactions between the central character and his partner, which were included in the "couple relationship" cluster. Twenty of 77 items in the memory questionnaire were not entered into analysis, as the agreement criterion was not reached.

For each content cluster, we examined proportion of recall and recognition attempts (Fig. 5A) (correct and incorrect answers and confidence levels were collapsed) and proportion of correct recall and correct HCR answers (Fig. 5B). Statistical analysis was performed for proportion of recall attempts (as recognition attempts measure was complementary) and for proportion of correct recall (analysis could not be performed on proportion of correct HCR, as several groups' data were uniform, i.e. 100% of subjects that answered questions in a given cluster using HCR, answered correctly).

We find that narrative elements and social interactions between characters are remembered best ("plot themes" questions elicited close to 100% accuracy in all ST groups). Questions about "jokes" and "details" elicit less recall attempts and answer accuracy declines more rapidly and drastically than for all other question clusters. Better performance of ST groups relative to LT groups is established for all content clusters, while comparison of time-dependent performance decline (i.e., curve slope) between content clusters approached significance only for correct recall performance.

Significant main effects of content and time-interval group were found (recall attempts: content $F_{(4,3224)} = 62.27$, $P < 0.0001$, time-interval group $F_{(4,44.13)} = 4.96$, $P < 0.002$; correct recall: content $F_{(4,874)} = 8.02$, $P < 0.0001$; time-interval group $F_{(4,75.54)} = 8.49$, $P < 0.0001$). The interaction between these effects (difference in performance between content clusters as a function of time-interval group, or comparison of slopes) was insignificant for both analyses, but approached significance for correct recall (recall attempts: $F_{(16,3224)} = 1.1$, $P < 0.35$; correct recall: $F_{(16,874)} = 1.52$, $P < 0.09$). In order to further explore main effects, pairwise contrasts were performed separately according to time-interval groups (Table 1) and according to content clusters (Table 2).

It is noteworthy that the division of LTM performance into ST and LT groups, introduced above on the basis of analysis of correct and incorrect answers, is also supported by the analysis of content-based correct recall answers, and is partially supported by analysis of content-based recall attempts. Subjects in the ST groups made significantly more recall attempts and more correct recall answers, per content cluster, than subjects in the LT groups. (Two exceptions are the "couple relationship" question cluster, which elicited a high proportion of correct recall answers throughout the tested span, and proportion of recall attempts of the 3-wk group that did not differ from both LT groups; Tables 1, 3.) While superior performance across content clusters is found

Table 1. Comparison of recall between time-interval groups

Time-interval groups contrasted	Recall attempts			Correct recall		
	DF	T	Significance	DF	T	Significance
3 h–3 mo	43.47	3.46	$P < 0.05$	49.55	3.68	$P < 0.01$
3 h–9 mo	43.75	3.20	$P < 0.05$	61.56	3.22	$P < 0.05$
1 wk–3 mo	43.23	3.01	$P < 0.05$	74.77	4.13	$P < 0.001$
1 wk–9 mo	43.54	2.77	$P < 0.08$	86.46	3.68	$P < 0.005$
3 wk–3 mo	46.49	1.70	$P < 0.7$	105	3.81	$P < 0.005$
3 wk–9 mo	46.4	1.51	$P < 0.8$	116.3	3.40	$P < 0.01$
3–9 mo	49.48	−0.06	$P < 1$	83.83	−0.08	$P = 1$

Time-interval groups are groups of participants tested at the indicated time after watching the movie. (DF) Degrees of freedom.

for the ST groups, different content clusters elicited significantly different performance profiles over time. This is illustrated by observing that the starting point of maximal value for forgetting curves varies between plots of different content clusters (Fig. 5A,B).

We should qualify our finding that memory performance can be differentiated according to content elements, as it seems that time since encoding is also an important factor in rendering this differentiation evident. Further examination of correct recall data, where interaction between content and group effects was found to only approach significance, was performed using post hoc comparisons. When examining content effects within time-interval groups, significant differences between content clusters were found in all but the 3-h and 1-wk groups (Table 3). When comparing time-interval group effects within content clusters, significant differences between groups were found in all but the "couple relationship' cluster (Table 3). As many comparisons were made, most pairwise contrasts did not survive correction of *P*-value calculation.

Manipulation of movie and questionnaire did not diminish memory

We further tested whether certain manipulations of the movie during the study session, or of the questionnaire in the test session, or both, will affect memory performance. A separate set of experimental groups all participated in the test session 3 wk after watching the movie, but were subjected to manipulated study or test material. Manipulations were either in the order of content material (scrambling the order of scenes in the movie or the questions in the test) or in perceptual attributes (eliminating color from the movie). One experimental group performed an interference protocol, in which subjects watched a different episode from the same sitcom at the beginning of the test session, and immediately afterward completed the original computerized questionnaire. No significant differences were found in performance (correct answers, collapsing confidence levels) between manipulation protocols and the original 3-wk group (unaltered movie and test, $n = 8$, 78.79 ± 2.02% correct; scrambled version of the movie followed by regular test, $n = 10$, 79.87 ± 2.38% correct; unaltered movie followed by scrambled test, $n = 6$, 73.81 ± 2.97% correct; scrambled movie followed by scrambled test, $n = 7$, 73.1 ± 2.76% correct; interference protocol, $n = 7$, 73.28 ± 2.39% correct; regular movie in black and white and regular test using black and white frames as visual cues, $n = 6$, 70.35 ± 2.99% correct; $F_{(5,40.25)} = 2.24$, $P < 0.07$). It is noteworthy that although some of the manipulations did show a trend for decreased performance, scrambling of the order of the scenes in the movie itself had no effect whatsoever. The potential implication of this finding to the encoding of the study material is discussed below.

Discussion

We describe a memory paradigm in which the study material is a 27-min narrative movie. This paradigm was intended to mimic aspects of "real-life" learning and memory under controlled experimental settings. We tested the memory once, in delays ranging from 3 h to 9 mo after the study session. The test targeted events that occur in the movie every ~20 sec. We found that details from the movie, which the participants watched only once without prior instruction to remember it, were remembered well over several months. Multiple performance measures indicate that long-term memory after hours-to-weeks is different from memory performance after several months. Recall answers, which we considered as the highest confidence answers, proved to be reliable measures of memory only for shorter durations, while HCR answers were highly reliable throughout the measured time span. Despite manipulation of movie and test materials, meant to disrupt narrative construction during encoding and/or retrieval, memory performance was unaffected. One possible explanation is that subjects were still able to successfully reconstruct the narrative from the scrambled segments. We further demonstrate that information content significantly influences memorability over time (though some time is needed for this effect to become evident). Finally, we use the unique resolution of our memory questionnaire to suggest that memory capacity for real-life information might be higher than previously estimated.

In clinical neurology and in cellular neurobiology, memory after 3 h is already long-term memory, and memory after 9 mo could be considered remote memory. Studies of human long-term memory and remote memory are abundant in the literature (for review, see Rubin and Wenzel 1996; Baddeley 1997; Moscovitch et al. 2006; Squire and Bayley 2007). These studies provide highly valuable data and insight on the performance of human memory over long periods, but differ in many aspects, including the nature of the study material and the memory system involved, the relevance of this material to "real life," and the potential of reproducibility of the protocol among participants and studies. In many cases, the need to use highly controlled and reproducible memoranda in laboratory settings dictates the use of study items such as individual word lists and verbal or visual paired-associates with only limited relevance to real-life (Neisser 1978). Retention in such artificial tasks is commonly tested at short intervals (for review, see Rubin and Wenzel 1996).

The present study is more akin to the category of studies that use more naturalistic material, which is more easily retained for long periods, but more difficult to control in laboratory settings. Such studies use a wide spectrum of types of information and conditions, and tax memory up to many years after encoding. They assess a wide spectrum of knowledge, ranging from semantic knowledge acquired intentionally (Tulving 1983; Bahrick 1984, 2000; Conway 1991), including public information such as famous faces or events acquired incidentally (McGehee 1937; Cohen and Squire 1981; Squire 1989; Reed and Squire 1998), to events witnessed and experienced (Bartlett 1932; Wells and Loftus 1984), and autobiographical memory (Rubin 1986; Wagenaar 1986; Conway and Holmes 2004).

In many of these studies, control over the encoding situation and the ability to reproduce it are limited or nonexistent. Attempts were made to overcome such shortcomings. An example is the diary method in autobiographical memory, in which the experimenter commits daily events to a diary and tests

Table 2. Comparison of recall between content-clusters

	Recall attempts		Correct recall	
Content-clusters	$T_{(3224)}$	Significance	$T_{(3224)}$	Significance
Plot Themes—Social Interactions	9.2	$P < 0.0001$	4.09	$P < 0.0005$
Plot Themes—Couple Relationship	7.79	$P < 0.0001$	−1.36	$P < 0.9$
Plot Themes—Jokes	9.94	$P < 0.0001$	−3.27	$P < 0.05$
Plot Themes—Minor Details	14.5	$P < 0.0001$	−5.23	$P < 0.0001$
Social Interaction—Couple Relationship	1.33	$P < 0.9$	1.54	$P < 0.8$
Social Interaction—Jokes	4.57	$P < 0.0001$	−0.86	$P < 1$
Social Interaction—Minor Details	7.53	$P < 0.0001$	−2.07	$P < 0.4$
Couple Relationship—Jokes	2.95	$P < 0.05$	1.80	$P < 0.6$
Couple Relationship—Minor Details	4.53	$P < 0.0001$	2.78	$P < 0.06$
Jokes—Minor Details	0.75	$P < 1$	0.46	$P = 1$

Content-clusters are clusters of questions that were grouped into predetermined content categories by independent raters as detailed in the text.

their recollections years later (Wagenaar 1986). Similarly, contemporary technology permits video recording of the study session. This improves the ability to verify the veridicality of recollection in the test session, but since episodes in real life are by definition unique, it doesn't solve the problem of reproducibility.

Movies provide an easily accessible solution to some of the aforementioned issues. The use of movies to tax memory can be traced to the early days of cinema (Boring 1916), but didn't catch on, with rare exceptions (Beckner et al. 2006). Narrative movies offer the advantage of simulating a continuous real-life episode that is highly reproducible. In this study, we have selected a 27-min movie that depicts rather ordinary life events without outstandingly funny, sad, or other emotionally arousing occurrences. The computerized test questionnaire was designed to sample the movie continuously, in segments of about 20 sec each, while assessing multiple manifestations of recollection, including cued recall, content recognition, and metamemory. We are unaware of previous investigations of the memory of a continuous episode that combined the resolution of performance and memory span of the present study using a readily reproducible study material.

On the robustness of memory

Some studies of semantic and autobiographical memory report highly robust remote memory. Tests of fact knowledge unveiled >80% correct responses over 30 yr, and in some tasks, e.g., famous names recognition, was >90% (Reed and Squire 1998). In another type of study, Wagenaar (1986) reports that using the diary method, recollection of 80% of the recorded events after 5 yr, and up to complete recollection provided proper cues were provided. However, most of us are exposed to salient public facts recurrently, over rather long periods, and autobiographical events might carry a saliency valence that reinforces their encoding. These considerations do not apply to the present study material. Why should the memory of details in a random, not aesthetically, literary, or conceptually unique movie, watched only once, linger for many months?

It is plausible to consider two corresponding points. First, humans are storytellers who weave mental narratives that fit schemata and support memory (Polkinghorne 1991). Indeed, people remember information much better when it is presented as narrative (Lichtenstein and Brewer 1980) and are prone to remember well the content of novels, plays, and movies (Rubin 1977; Cohen 1996; Neisser 1998). The study of memory for short stories indicates, however, that people commonly do not remember a narrative verbatim, but rather reconstruct it according to experience and cultural assumptions using schemes as frameworks for retrieval (Bartlett 1932; Mandler and Johnson 1977; Neisser 1998; Fivush and Nelson 2004). This tendency may in fact increase erroneous memory for fine details.

Most studies of narrative memory so far have used a single, temporally proximal retrieval test. Exceptions include a recognition test of verbatim and paraphrased sentences from a short narrative over 4 d (Kintsch et al. 1990), and a cross-sectional study of long-term memory for a Dickens novel (Stanhope et al. 1993). The latter study used free recall to probe memory for character names and their roles in the plot, finding 45% correct recall of both names and roles after 3 mo, maintaining 35% of roles and 20% of names after more than 3 yr. For comparison, in our study, the experimental group tested 3 mo after watching the narrative movie correctly answered 70% of overall questions (Fig. 2A), and correct recall answers to questions by content were found for >70% of plot theme questions, 45% of minor detail questions, 62% and 73% of social interactions, and couple relationship, respectively (Fig. 5B). Although testing methods vary considerably between these two studies, it is plausible to conclude that a retention interval of 3 mo enables better recollection of narrative when it was seen in a movie than when it was read in a novel.

A second consideration is the notion that in order to learn something, one must already know a lot (Charniak and McDermott 1985). There is evidence that acquisition of new knowledge is conditioned by previously acquired relevant knowledge (Alba and Hasher 1983; Ericsson and Lehmann 1996). Although it is arguable whether the young adults who participated in our experiments have developed an expertise in learning from watching sitcoms, it can be claimed that we are all experts in incidental encoding of socially relevant information from ongoing continuous multimodal stimuli. This implies that observers of the movie might have already had lots of knowledge into which the new information integrated well.

Changes in memory performance over time

Multiple measures in our study combine to indicate that more recent memory differs from more remote memory. Hence, significant differences between ST and LT groups are revealed from analysis of percent of correct answers (Fig. 2A), recall attempts

Table 3. Time-content interactions

	Statistic	Significance
Time-interval group effects within content cluster		
Plot Themes	$F_{(4,276.5)} = 7.51$	$P < 0.0001$
Social Interactions	$F_{(4,121.3)} = 6.48$	$P < 0.0001$
Couple Relationship	$F_{(4,874)} = 0.8$	$P < 0.6$
Jokes	$F_{(4,874)} = 2.82$	$P < 0.05$
Minor Details	$F_{(4,406.7)} = 3.92$	$P < 0.005$
Content cluster effects within time-interval group		
3 h	$F_{(4,874)} = 1.27$	$P < 0.3$
1 wk	$F_{(4,874)} = 1.82$	$P < 0.2$
3 wk	$F_{(4,874)} = 4.41$	$P < 0.005$
3 mo	$F_{(4,874)} = 2.72$	$P < 0.05$
9 mo	$F_{(4,874)} = 2.76$	$P < 0.05$

Time-interval groups and content-clusters are as in Tables 1 and 2, respectively.

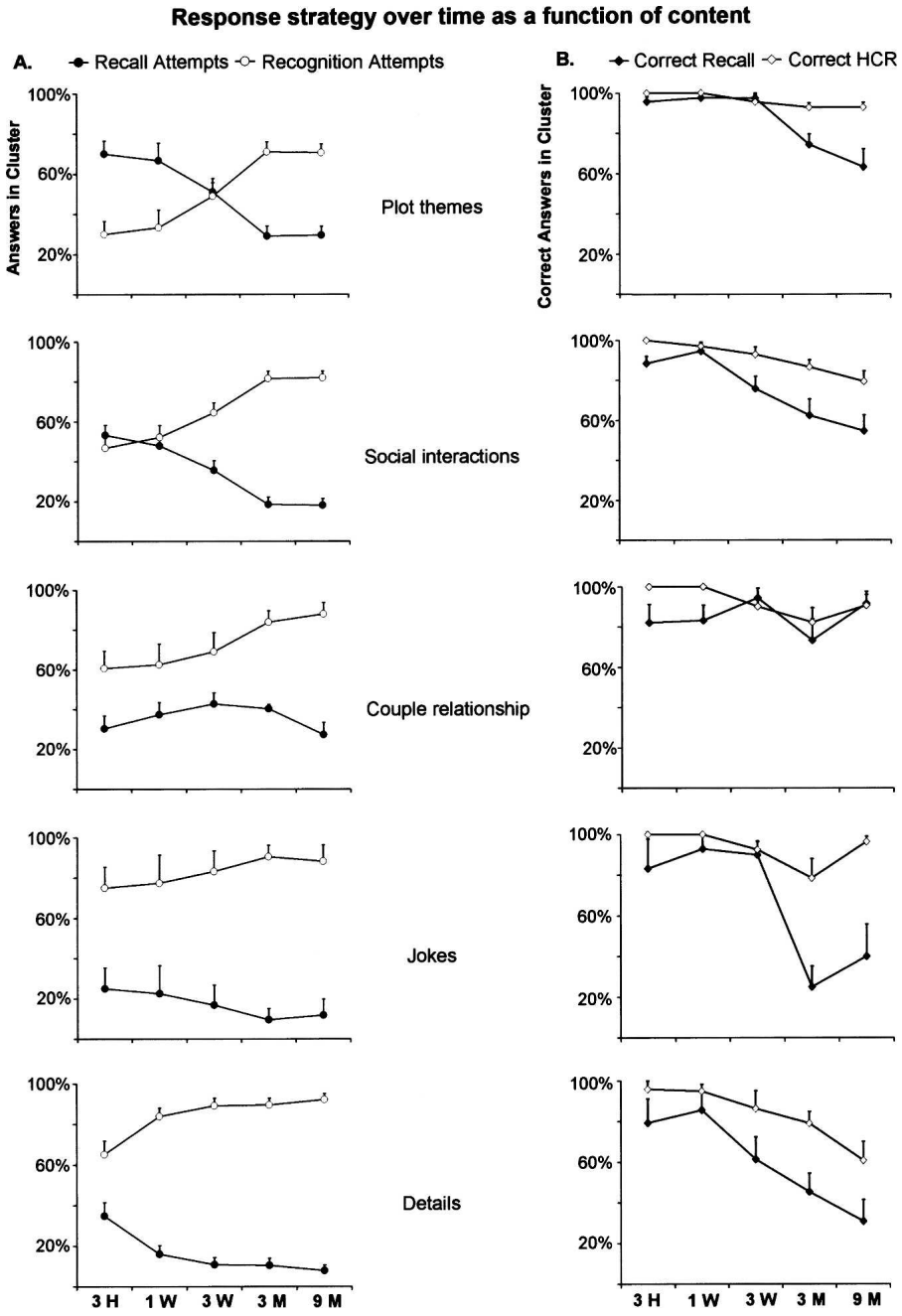

Figure 5. Response strategy over time as a function of content. Questions were grouped into clusters based on 66% agreement between raters. Clusters were dubbed "plot themes," "social interactions," "couple relationship," "jokes," and "minor details" (13, 19, 6, 5, and 14 questions, respectively). (*A*) Recall and recognition attempts as a function of content. Each panel describes average percent of answers (correct + incorrect) given using recall (●) and recognition (○) for each content cluster. Recognition curve includes all confidence levels and, therefore, complements the recall curve. Recall curves of different clusters are significantly different and show significant decline over time. (For statistics, see Results.) (*B*) Correct recall and correct high-confidence recognition (HCR) answers as a function of content. Each panel describes proportion of correct recall (♦) and correct HCR (◊) answers to questions in each cluster. Proportions were calculated from answers made with each confidence level, per time-interval group. Correct recall curves of different clusters are significantly different and show significant decline over time. (For statistics, see Results.) Values are mean ± SEM.

(Fig. 2B), overall performance by level of confidence (Fig. 3A), correct/incorrect answers by level of confidence (Figs. 3B, 4), and from analysis of content-based recall answers (Fig. 5A,B). Using all of these measures, we find no significant performance differences between groups within ST subdivision or within LT subdivision, but significantly better performance of ST vs. LT groups. This corroborates the notion of a nonmonolithic nature of long-term memory, while leaving open the question of the kinetics of decline in performance between retention intervals of weeks to months. These findings could be interpreted in line with systems-consolidation hypotheses, which posit that with the passage of time, declarative information becomes differentially reliant on certain brain circuits (for variants of systems consolidation models, see McClelland et al. 1995; Squire and Alvarez 1995; Nadel and Moscovitch 1997; Dudai 2004; Moscovitch et al. 2005). We must await, however, brain studies to support this hypothesis.

Effect of content

An "ecological" view of memorability takes into consideration the type of items to be recalled (Kintsch et al. 1990; Koriat et al. 2000). The complex nature of a movie stimulus and the high resolution we used in tapping into the memory of this material allowed us to re-examine performance by dividing the questionnaire, which was originally composed according to temporal segmentation, into clusters of questions sharing similar content. This content-based analysis revealed significantly better performance for questions relating to plot themes and social interactions relative to jokes and details (Fig 5). The significant difference between content clusters (Table 2) reaffirms an attribute of the "correspondence approach," which is concerned primarily with reliability and accuracy of memory rather than with its magnitude. This approach stipulates that not all memory items are identical, and that it is instructive to note which items are remembered and which are misremembered (Koriat and Goldsmith 1996). Our findings indicate not only that some content elements are remembered better, but also that memory performance across a time span of days, weeks, and months can be differentiated according to the content of the memory trace. This replicates findings from studies of long-term retention of information originating in real-life experiences, which suggest differential forgetting rates for different knowledge aspects (Bahrick 1984; Conway 1991).

The highest performance measures are seen for questions in the "plot themes" cluster; "jokes" and "minor details" elicit the lowest proportion of recall attempts and the steepest decline in proportion of correct recall (Fig. 5A,B). Controlled lab experiments using sentence and narrative story memory have yielded similar

results, concluding that main plot themes are remembered better and for longer durations than verbatim quotes or other details (Kintsch et al. 1990; Koriat et al. 2000). Social information was also well recalled throughout the measured time span, recapitulating findings from personality and social psychology that suggest that social monitoring, motivated by a powerful need to belong to one's social group, enhances processing and memory of such information (Gardner et al. 2000). Our finding that jokes were not remembered well is in contrast with previous findings that humor aids memory (Schmidt 1994). It could be argued, however, that a punchline in a sitcom is diluted with other humorous occurrences, and therefore is not as salient as in the out-of-sitcom world. Finally, a caveat is in place: differential rates of forgetting of various content domains may reflect not only differences in maintenance of these content categories, but also possible differences in degree of original learning (Bahrick 2000).

Mode of retrieval and metamemory

A feature of the protocol described here is the interactive nature of the test session, which allows participants to make a choice between two retrieval modes for each questionnaire item separately. This flexible design allows us to follow metamemory dynamically, while also providing better correspondence to the use of memory in real life, in which subjects adjust retrieval strategies based on need and occasion.

Stability in use and accuracy of high confidence in the "recognition mode" (HCR) is noteworthy, especially when compared with cued recall. Some previous studies show that recall performance declined more rapidly than recognition performance (Bahrick et al. 1975). Although our use of "recognition" refers to content and not to the exact target item (and might actually fit better the meaning of recognition in real-life, as in recognizing a person in different modalities or age), the aforementioned trend was observed in our study as well: while proportions of recall attempts and correct recall answers declined sharply, particularly in LT groups, HCR attempts did not change significantly, and correct HCR answers decline modestly (98% correct after 3 h, >80% correct after 9 mo; Fig. 3). The stable performance using HCR was also demonstrated in content-based analysis, as proportion of correct HCR in several ST groups was 100%, preventing statistical analysis (Fig. 5B).

While cued recall was considered to reflect higher confidence than recognition measures (Hart 1965), it seems that this is true in our study only for shorter retention intervals. A significant decline in recall performance was observed when moving from ST to LT groups (Fig. 3), particularly for decontextualized information like jokes and minor details (Fig. 5B). Recall was very much used when subjects attempted to recollect salient details that constitute the gist of the experienced or witnessed event, like the main plot themes (Fig. 5A). This supports the use of introspective metamemory measurements as informative measures for the strength of memory traces for real-life like events, similarly to their use in memory tests employing verbal material.

On the density and capacity of memory

Attempts to gauge the capacity of human memory are numerous, but involve many uncertainties (Dudai 1997). We exploited the unique properties of the present protocol, particularly the availability of memory performance scores that sample events three times per minute in an extended episode, to speculate about the density of retrievable long-term memory. The inherent limitations in our estimate notwithstanding, we reach an order of magnitude of approximately an event per minute recalled after months. The number game can become overenticing, but in the absence of more robust estimates, it is tempting to proceed one arithmetic step further: 16 h encoding per day yields, with that order of magnitude, a potential capacity of 10^3 items per day, or an order of magnitude of 10^7 per lifetime. This is still three orders of magnitude smaller than the estimated number of percepts a human brain may be able to acquire in a lifetime based on psychophysics, but higher than some earlier estimates for episodic memory, though none of the earlier estimates was able to measure the density of retrievable memory with the resolution used here (Dudai 1997). Since we clearly cannot claim that our test saturates memory capacity, the eternal question of whether we are capable of encoding all that we perceive (Burnham 1889; Loftus and Loftus 1980; Dudai 2002) remains open, to the delight of future players of similar number games.

Erroneous memory

Manipulations of question wording and other types of suggestions have been shown to cause subjects to endorse incorrect answers with high levels of confidence (Loftus 1993, 2005). In our study, while the proportion of incorrect recall answers seemed to rise more steeply than the proportion of incorrect HCR answers, complementary analysis of the average number of incorrect answers by confidence reveals that this was caused by a general decline in total proportion of recall answers made (Fig. 4). It is noteworthy, however, that whether participants were tested a few hours, a few weeks, or a few months after watching the movie, there were always a few questions that they believed they knew the answer to sufficiently to take the recall challenge, when in fact they did not (Fig. 4B, incorrect recall). We could not identify a particular subset of items in our test questionnaire, the response to which could explain this observation. That erroneous recall is evident already shortly after a seemingly nondramatic event, only reinforces the skepticism concerning the reliability of even utterly naive, nonbiased eyewitnesses (Loftus 1996).

A possible explanation for the occurrence of erroneous recall already shortly after watching the movie is that events perceived are integrated erroneously into previous knowledge schemas (Bartlett 1932). Integration of new information into existing schemas can occur rapidly in memory consolidation (Tse et al. 2007). A related possibility, which might also contribute to the overall decline in correct responses over time, is that either the movie narrative or the test questionnaire, or both, promoted confusion of contextual cues. Hence, while every item in our questionnaire contains visual and semantic cues that were selected to pertain to a specific scene, similarities or mental reconstruction of movie narrative content might have created contextual cues for other events. This could have engendered false answers with a high level of confidence, an effect reminiscent of daily life occurrences (Schwartz et al. 2005). It is noteworthy that Bahrick et al. (1975) interpreted intrusion errors in free recall over time as evidence for loss of context. The influence of context, in the larger sense of emotional involvement and real-life motivations, was also previously demonstrated to affect perception and recollection in the classic study by Hastorf and Cantril (1954), which examined memory for a controversial football game, and to which the present paper owes its title. Finally, particularly since the movie depicts real-life types of events, buildup of interference over time, because retrieval cues become less effective as they become associated with additional items (for review, see Wixted 2004), should also be considered as potentially contributing to the decrease in correct responses over time.

Immunity to manipulations of study material

We attempted to disrupt the creation of narrative context by scrambling the movie scenes (but not shots within the scenes),

making it possibly harder for subjects to piece together the narrative of the movie (Kintsch et al. 1977). We also disrupted the order of the questions in the questionnaire, such that it would not support gradual construction of the narrative while answering the questions. These manipulations were based on the premise that people are sensitive to temporal order information (Alba and Hasher 1983; Zacks et al. 1984). A different kind of manipulation involved obstructing perceptual information such as color, as it has been demonstrated that color benefits both encoding and recognition of natural scene images (Gegenfurtner and Rieger 2000; Wichmann et al. 2002). We therefore speculated that using black and white study/test materials would cause a decline in memory performance when compared with the regular protocol. Yet another kind of manipulation attempted retroactive interference of retrieval by showing subjects a different episode from the same sitcom just before answering the questionnaire (Wixted 2004). It is noteworthy that none of these manipulations significantly altered memory performance, though trends for decline in performance in some groups were noted. Of particular interest is the total lack of effect in memory performance in the scrambled scene movie. While it is not trivial to explicitly measure how and when subjects reconstructed the narrative when shown a scrambled version, one possibility of explaining this finding is postulating that memory was dependent on deeper levels of processing of the movie material, such as the narrative construction, possibly both on-the-go in real time and in subsequent consolidation. Such reconstruction might actually be promoted by a scrambled scene version of the movie.

In conclusion, we describe an experimental paradigm that permits fine analysis of multiple facets of the memory of a controllable and reproducible continuous episode. This paradigm can be used to investigate specific questions about the influence of cognitive factors (such as emotional saliency, theory of mind capacity, or gender differences) on the formation and long-term maintenance of memories that are encoded incidentally in a manner resembling real-life situations. This movie memory protocol is useful not only for the analysis of behavioral performance, but is particularly suitable for studying brain substrates and processes of real-life memory using functional brain imaging.

Materials and Methods

Participants

A total of 107 participants (73 female, mean age 25.4 ± 3.8 yr) took part in the study. Participants were recruited from a nearby university campus and tested at the Weizmann Institute of Science. Exclusion criteria were insufficient knowledge of the English language and familiarity with the TV sitcom used. Participants were remunerated for their time.

Study material and memory test

The study material was a 27-min episode from an English-language television sitcom that portrays a character's real-life-like actions taking place in a big city (*Curb Your Enthusiasm*, by Larry David, Home Box Office, Inc.). The episode used (Season 1, Chapter 7) contained ordinary types of events (e.g. a dinner party, arguments with friends).

The memory test was administered via a computerized interactive questionnaire (in-house software). Seventy-seven questions targeting memory for events across the episode were composed. They were chosen to cover the content of the entire episode, at an approximate rate of one question per 20 sec of movie time. A special emphasis was placed on including questions that targeted distinct events (i.e., taking place in an identifiable time segment), and to minimize the subject's ability to provide the correct answer by using information obtained either before or after the targeted segment of the movie. Unless otherwise indicated, questions were presented in the order the respective events occurred in the movie.

Each question was accompanied by the presentation of a still-frame from the movie that served as a visual cue. Subjects were given two options for answering each question: they were instructed to type a free-text answer if they were confident they could recall the answer (recall mode, i.e., cued recall; Fig. 1A), or, if they were not confident in their response, they completed a two-alternative forced choice which we refer to as recognition test. Subjects were also asked to report their subjective confidence for the recognition portion (highly confident, fairly confident, or guessing; Fig. 1B).[6]

Experimental protocols

Participants took part in two experimental sessions, study and test, that took place in a quiet, windowless room containing a desktop computer with a 19" CRT flat screen. In the study session, participants read instructions informing them that they would participate in "an experiment in the context of investigating human reaction to movies," and were asked to "watch the following movie, try to focus on the movie, and remain concentrated throughout." They then watched the episode individually on the screen, using headphones to optimize comprehension of spoken dialog. Subjects were informed that they might be called upon for an additional session at an unspecified time afterward within the context of the experiment, but were not informed of an impending memory questionnaire and were not asked to memorize movie content.

During the test session, participants viewed instructions on how to answer the computerized memory questionnaire, and completed a training demo. It was emphasized that participants should attempt to recall (i.e., enter a free-text response) only if they felt very confident that they knew the correct answer. Otherwise, they were instructed to choose the forced-choice recognition test. After completing the full questionnaire, participants were debriefed and were asked to rate their level of understanding of the movie and questions, and declare whether or not they had watched the sitcom on television during the interval between the study and test sessions. Four participants that did watch episodes of that sitcom in between the study and the test session were excluded from the experimental groups and two additional participants were excluded as they did not make themselves available for the test session. Both study and test sessions were completed at a similar time of day (most within range of ±3 h); all sessions took place between late morning and early evening (10 am to 8 pm).

Each experimental group was tested at varying post-encoding time intervals: 3 h (n = 8), 1 wk (n = 8), 3 wk (n = 12), 3 mo (n = 17), and 9 mo (n = 12). The control group (No-Movie, n = 8) did not watch the movie, and subjects in this group were encouraged to complete the questionnaire by inferring information from questions and visual cues or by guessing. The larger number of participants in the longer-time intervals was not intentionally included in the design of the study, but rather stems from precautionary recruitment of more participants for these time points, suspecting that fewer participants will return for testing weeks or months after the study session. This did not happen.

In addition, other experimental groups were tested 3 wk

[6]A caveat is appropriate concerning the use of the term recognition to designate the second reply mode in our test. By definition, recognition is the judgment of previous occurrence of the on-line item (Dudai 2002). In our protocol, the stimuli in the movie are audiovisual, whereas the test provides two written reply options. Hence, this mode of reply might also be considered as option-restricted or option-guided cued recall. However, since we considered the alternative choices as providing the opportunity to recognize the content of the specific event in the movie, and since the term recognition is sometimes generalized in the human memory literature to mean content recognition rather than exact modality-specific recognition, we selected the term recognition to distinguish the more extensively guided mode of reply from the initial recall opportunity.

after watching the movie, but using different protocols than the regular movie and regular test described above. These experimental groups were: (1) "Scrambled Movie," where scenes from the movie were intermixed in order to yield a scrambled version of the movie, and regular test ($n = 10$); (2) "Scrambled Test," where a regular movie was followed by altering the order of questions in the questionnaire to yield a scrambled test ($n = 6$); (3) "Scrambled Movie and Test" ($n = 7$); (4) "Interference protocol," where a regular movie and test were used, but a second narrative movie was presented to subjects immediately prior to regular test ($n = 7$), and (5) "Black and White," a black and white version of the regular movie was shown, and regular test was used with visual cues in black and white ($n = 6$).

Recall answers were coded manually as correct or incorrect (paraphrased answers were considered correct as their content was veridical), while results from the recognition portion were coded automatically by in-house software. Thus, two different ways of coding the memory outcome were used. The first was binary, with answers coded as correct or incorrect. The second was a four-step rating of level of confidence, irrespective of accuracy, the highest level being recall, followed by high-confidence recognition (HCR), low-confidence recognition (LCR), and guessing (Fig. 1C).

Measure for memory density

We devised a measure for estimating the quantity of memory stored after watching the movie in order to enable comparison between different time intervals. We started by calculating the chance to answer correctly (CAC) separately for every response mode, confidence level, and time-interval group. CAC was first calculated per subject, dividing the number of correct answers by the overall number of answers made in each confidence level ($N_{\text{correct per level}}/N_{\text{overall per level}}$). Next, a group average was calculated: CAC using recall was high for ST groups (hours-to-weeks groups, termed Shorter-Time-Interval [ST], see Results and Discussion) (79%–91%) but dropped to 51%–54% in LT groups (months groups, termed Longer-Time-Interval [LT]); CAC using HCR ranged from >92% for ST groups, to >83% for LT groups; CAC using LCR was 66%–79% for ST groups; finally, CAC using guesses was around 50% for all time-interval groups (range 49%–55%). We then normalized CAC for recognition measures (HCR, LCR, and Guess answers), by deducting 50% from the group CAC, given that recognition answers required a choice between two alternatives, so 50% of correct answers could be attributed to chance performance. We considered chance-performance level for recall answers to be 0, and so CAC for recall was not normalized.

In addition to the data set of CAC, we also generated a data set of the proportions of overall answers made at each confidence level, per time-group (illustrated in Fig 3A). Given these two data sets, we next multiplied for each time-group and each confidence level the proportion of answers by the normalized CAC (as explained above, CAC for recall was not normalized). For example, the 1-wk group answered 33.7% of questions using HCR; CAC using HCR was 96%, normalized CAC using HCR in this group were therefore 96% − 50% = 46%. Multiplying 33.7% and 46%, we get 15.5% correct answers using HCR. We then summed, for each time-interval group, the resulting proportions of correct answers across confidence levels. The contribution of the information contained in the questionnaire itself, to performance, was calculated as 2%, using the above procedure on performance of the No-Movie control group, which did not see the movie. Therefore, 2% were further deducted from the summed proportions.

Statistics

The data in this study are hierarchical (multilevel), as observations are answers of questions nested within subjects, and the dependent variables are binary, as answers can be correct or incorrect. In order to determine statistical significance we used logistic regressions, which belong to the family of Generalized Linear Models (GLM) (McCullagh and Nelder 1989; Hosmer and Lemeshow 2000). A "mixed model" was used, as it is a common way of dealing with hierarchical data, by accounting for the random subject effect in addition to the other fixed experimental conditions, such as level of confidence. Therefore, the regression models applied were logistic regressions for repeated measures, as implemented by the Generalized Linear Mixed Models (GLIMMIX) procedure in SAS. Data variance was not equivalent across subject groups; therefore, degrees of freedom were estimated using Welch-Satterthwaite approximation, leading to non-integer values. Follow-up contrasts used Sidak correction for multiple comparisons (Sidak 1967).

Acknowledgments

We thank Efrat Furst, Matthieu Guitton, Sharon Haramati, Daniel Levy, Kelly Ludmer, and Avi Mendelsohn for valuable discussion and comments. This work was supported by The Weizmann Institute-NYU Collaborative Fund in the Neurosciences (Y.D. and L.D.), The Seaver Foundation (to L.D.), and a HFSP long-term fellowship (U.H.). Part of the work on this manuscript was conducted while Y.D. visited NYU as the Albert and Blanche Willner Family Global Distinguished Professor of Neuroscience at the Center for Neural Science.

References

Alba, J.W. and Hasher, L. 1983. Is memory schematic. *Psychol. Bull.* **93:** 203–231.

Baddeley, A. 1997. *Human memory. Theory and practice*. Psychology Press, Hove, UK.

Bahrick, H.P. 1984. Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *J. Exp. Psychol. Gen.* **113:** 1–29.

Bahrick, H.P. 2000. Long-term maintenance of knowledge. In *Oxford handbook of memory*. (eds. E. Tulving and F. Craik), pp. 347–362. Oxford University Press, New York.

Bahrick, H.P., Bahrick, P.O., and Wittlinger, R.P. 1975. Fifty years of memory for names and faces: A cross-sectional approach. *J. Exp. Psychol. Gen.* **104:** 54–75.

Bartlett, F.C. 1932. *Remembering: A study in experimental and social psychology*. Cambridge University Press, Cambridge, UK.

Beckner, V.E., Tucker, D.M., Delville, Y., and Mohr, D.C. 2006. Stress facilitates consolidation of verbal memory for a film but does not affect retrieval. *Behav. Neurosci.* **120:** 518–527.

Boring, E.G. 1916. Capacity to report upon moving pictures as conditioned by sex and age. *J. Am. Inst. Crim. Law Criminol.* **6:** 820–834.

Buckner, R.L., Logan, J., Donaldson, D.J., and Wheeler, M.E. 2000. Cognitive neuroscience of episodic memory encoding. *Acta Psychol.* **105:** 127–139.

Burnham, W.H. 1889. Memory, historically and experimentally considered. *Am. J. Psychol.* **2:** 39–90.

Charniak, E. and McDermott, D. 1985. *Introduction to artificial intelligence*. Addison Wesley, Reading, MA.

Cohen, G. 1996. *Memory in the real world*, 2d ed. Psychology Press, Hove, UK.

Cohen, N.J. and Squire, L.R. 1981. Retrograde amnesia and remote memory impairment. *Neuropsychologia* **19:** 337–356.

Conway, M.A. 1991. On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *J. Exp. Psychol. Gen.* **120:** 358–372.

Conway, M.A. and Holmes, A. 2004. Psychosocial stages and the accessibility of autobiographical memories across the life cycle. *J. Pers.* **72:** 461–480.

Dudai, Y. 1997. How big is human memory, or, on being just useful enough. *Learn. Mem.* **3:** 341–365.

Dudai, Y. 2002. *Memory from A To Z. Keywords, concepts and beyond*. Oxford University Press, Oxford, UK.

Dudai, Y. 2004. The neurobiology of consolidations, or, how stable is the engram? *Annu. Rev. Psychol.* **55:** 51–86.

Eisenstein, S. 1969. *The film sense*. Harcourt Brace & Co, New York.

Ericsson, K.A. and Lehmann, A.C. 1996. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annu. Rev. Psychol.* **47:** 273–305.

Fivush, R. and Nelson, K. 2004. Culture and language in the emergence of autobiographical memory. *Psychol. Sci.* **15:** 573–577.

Gardner, W.L., Pickett, C.L., and Brewer, M.B. 2000. Social exclusion and selective memory: How the need to belong influences memory for social events. *Pers. Soc. Psychol. Bull.* **26:** 486–496.

Gegenfurtner, K.R. and Rieger, J. 2000. Sensory and cognitive contributions of color to the recognition of natural scenes. *Curr.*

Biol. **10:** 805–808.
Hart, J.T. 1965. Memory and the feeling-of-knowing experience. *J. Educ. Psychol.* **56:** 208–216.
Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. 2004. Intersubject synchronization of cortical activity during natural vision. *Science* **303:** 1634–1640.
Hastorf, A. and Cantril, H. 1954. They saw a game: A case study. *J. Abnorm. Psychol.* **49:** 129–134.
Hosmer, D.W. and Lemeshow, S. 2000. *Applied logistic regression*, 2d ed. John Wiley & Sons Inc., New York.
Kintsch, W., Mandel, T.S., and Kozminsky, E. 1977. Summarizing scrambled stories. *Mem. Cognit.* **5:** 547–552.
Kintsch, W., Welsch, S., Schmalhofer, F., and Zimny, S. 1990. Sentence memory: A theoretical analysis. *J. Mem. Lang.* **29:** 133–159.
Koriat, A. and Goldsmith, M. 1996. Monitoring and control processes in the strategic regulation of memory accuracy. *Psychol. Rev.* **103:** 490–517.
Koriat, A., Goldsmith, M., and Pansky, A. 2000. Toward a psychology of memory accuracy. *Annu. Rev. Psychol.* **51:** 481–537.
Lichtenstein, E.H. and Brewer, W.F. 1980. Memory for goal-directed events. *Cognit. Psychol.* **2:** 412–445.
Loftus, E.F. 1993. Made in memory: Distortions in recollection after misleading information. In *The psychology of learning and motivation: Advances in theory and research* (ed. D.L. Medin), pp. 187–215. Academic Press, New York.
Loftus, E.F. 1996. Memory distortion and false memory creation. *Bull. Am. Acad. Psychiatry Law* **24:** 281–295.
Loftus, E.F. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learn. Mem.* **12:** 361–366.
Loftus, E.F. and Loftus, G.R. 1980. On the permanence of stored information in the human brain. *Am. Psychol.* **35:** 409–420.
Mandler, J.M. and Johnson, N.S. 1977. Remembrance of things parsed: Story structure and recall. *Cognit. Psychol.* **9:** 111–151.
McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102:** 419–457.
McCullagh, P. and Nelder, J.A. 1989. *Generalized linear models*, 2d ed. Chapman and Hall, London, UK.
McGehee, F. 1937. The reliability of the identification of the human voice. *J. Gen. Psychol.* **17:** 249–271.
Morin, E. 2005. *The cinema, or the imaginary man*. University of Minnesota Press, Minneapolis.
Moscovitch, M., Rosenbaum, R.S., Gilboa, A., Addis, D.R., Westmacott, R., Grady, C., McAndrews, M.P., Levine, B., Black, S., Winocur, G., et al. 2005. Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *J. Anat.* **207:** 35–66.
Moscovitch, M., Nadel, L., Wincour, G., Gilboa, A., and Rosenbaum, R.S. 2006. The cognitive neuroscience of remote episodic, semantic and spatial memory. *Curr. Opin. Neurobiol.* **16:** 179–190.
Nadel, L. and Moscovitch, M. 1997. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7:** 217–227.
Neisser, U. 1978. Memory: What are the important questions? In *Practical aspects of memory* (eds. M.M. Grunberg et al.), pp. 3–24. Academy Press, London, UK.
Neisser, U. 1998. Stories, selves and schemata: A review of ecological findings. In *Theories of memory II* (eds. M.A. Conway et al.), pp. 171–186. Psychology Press, Hove, UK.
Polkinghorne, D.E. 1991. Narrative and self-concept. *J. Narrative and Life-History* **1:** 135–153.
Reed, J.M. and Squire, L.R. 1998. Retrograde amnesia for facts and events: Findings from four new cases. *J. Neurosci.* **18:** 3943–3954.
Rubin, D.C. 1977. Very long-term memory for prose and verse. *J. Verb. Learn. and Verb. Behav.* **16:** 611–621.
Rubin, D.C., ed. 1986. *Autobiographical memory*. Cambridge University Press, Cambridge, UK.
Rubin, D.C. and Wenzel, A.E. 1996. One hundred years of forgetting: A quantitative description of retention. *Psychol. Rev.* **103:** 734–760.
Schmidt, S.R. 1994. Effects of humor on sentences memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **20:** 953–967.
Schwartz, G., Howard, M.W., Jing, B., and Kahana, M.J. 2005. Shadows of the past: Temporal retrieval effects in recognition memory. *Psychol. Sci.* **16:** 898–904.
Sidak, Z. 1967. Rectangular confidence regions for means of multivariate normal distributions. *J. Am. Stat. Assoc.* **62:** 626–633.
Squire, L.R. 1989. On the course of forgetting in very long-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **15:** 241–245.
Squire, L.R. and Alvarez, P. 1995. Retrograde amnesia and memory consolidation: A neurobiological perspective. *Curr. Biol.* **5:** 169–177.
Squire, L.R. and Bayley, P.J. 2007. The neuroscience of remote memory. *Curr. Opin. Neurobiol.* **17:** 1–12.
Stanhope, S., Cohen, G., and Conway, M.A. 1993. Very long-term retention of a novel. *Appl. Cogn. Psychol.* **7:** 239–256.
Suddendorf, T. and Busby, J. 2005. Making decisions with the future in mind: Developmental and comparative identification of mental time travel. *Learn. Motiv.* **36:** 110–125.
Tse, D., Langston, R.F., Kakeyama, M.M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., and Morris, R.G.M. 2007. Schemas and memory consolidation. *Science* **316:** 76–82.
Tulving, E. 1983. *Elements of episodic memory*. Oxford University Press, Oxford, UK.
Tulving, E. 2002. Episodic memory: From mind to brain. *Annu. Rev. Psychol.* **53:** 1–25.
Wagenaar, W.A. 1986. My memory: A study of autobiographical memory over six years. *Cognit. Psychol.* **18:** 225–252.
Wells, G.L. and Loftus, E.F., eds. 1984. *Eyewitness testimony. Psychological perspectives*. Cambridge University Press, New York.
Wichmann, F.A., Sharpe, L.T., and Gegenfurtner, K.R. 2002. The contributions of color to recognition memory for natural scenes. *J. Exp. Psychol. Learn. Mem. Cogn.* **28:** 509–520.
Winocur, G. and Weiskrantz, L. 1976. An investigation of paired-associate learning in amnesic patients. *Neuropsychologia* **14:** 97–110.
Wixted, J.T. 2004. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.* **55:** 235–269.
Zacks, R.T., Hasher, L., Alba, J.W., Sanft, H., and Rose, K.C. 1984. Is temporal order encoded automatically? *Mem. Cognit.* **12:** 387–394.

Received February 7, 2007; accepted in revised form May 1, 2007.