Research Report

Neural Evidence That Vivid Imagining Can Lead to False Remembering

Brian Gonsalves,[1] Paul J. Reber,[1,2,3] Darren R. Gitelman,[2,3,4] Todd B. Parrish,[2,3,5] M.-Marsel Mesulam,[2,3,4] and Ken A. Paller[1,2,3]

[1]Department of Psychology and [2]Institute for Neuroscience, Northwestern University, and [3]Cognitive Neurology and Alzheimer's Disease Center, [4]Department of Neurology, and [5]Department of Radiology, Northwestern University Medical School

ABSTRACT—*The imperfect nature of memory is highlighted by the regularity with which people fail to remember, or worse, remember something that never happened. We investigated the formation of a particular type of erroneous memory by monitoring brain activity using functional magnetic resonance imaging during the presentation of words and photos. Participants generated a visual image of a common object in response to each word. Subsequently, they sometimes claimed to have seen photos of specific objects they had imagined but not actually seen. In precuneus and inferior parietal regions of the cerebral cortex, activations in response to words were greater when participants subsequently claimed to have seen the corresponding object than when a false memory for that object was not subsequently produced. These findings indicate that brain activity reflecting the engagement of visual imagery can lead to falsely remembering something that was only imagined.*

Neuroscientific studies of false remembering form part of the broader research agenda of understanding how memories are formed in the brain, stored in an enduring way, and retrieved after some time has elapsed. The experiences, thoughts, and facts that people remember are normally not verbatim replays of the past, but rather are reconstructions based on a limited amount of information that is stored. Furthermore, not all information initially encoded is available at the time of retrieval, as memories tend to fade or become distorted with time (Bartlett, 1932; Roediger & McDermott, 2000; Schacter, 1995). The study of neural and cognitive processes responsible for the formation of memories in the human brain holds promise for elucidating why information can sometimes be remembered but other times is forgotten or distorted.

Address correspondence to Brian Gonsalves, Department of Psychology, Jordan Hall, Building 420, Stanford University, Stanford, CA 94305; e-mail: bgon@psych.stanford.edu.

Cognitive neuroscience techniques have been used to address these issues by examining brain activity at encoding as a function of later remembering or forgetting on an item-by-item basis (Paller & Wagner, 2002). This type of analysis, a *subsequent-memory*, or *Dm*, analysis, has been applied in numerous studies using event-related potentials, or ERPs (e.g., Fabiani, Karis, & Donchin, 1986; Paller, Kutas, & Mayes, 1987; Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980). Typically, a more positive ERP is elicited at encoding by items that are later accurately remembered than by those that are forgotten. Furthermore, studies using event-related functional magnetic resonance imaging (fMRI) have revealed greater left inferior prefrontal cortex activity during encoding of words subsequently remembered compared with those subsequently forgotten (Davachi, Maril, & Wagner, 2001; Henson, Rugg, Shallice, Josephs, & Dolan, 1999; Kirchhoff, Wagner, Maril, & Stern, 2000; Otten, Henson, & Rugg, 2001; Reber et al., 2002; Wagner et al., 1998). This left frontal Dm effect has been interpreted as an indication of greater phonological or semantic processing for words that are later remembered compared with words that are later forgotten, particularly when people are intentionally attempting to commit information to memory.

Sometimes, however, cognitive processes engaged at encoding lead to inaccurate remembering later. For example, so-called reality-monitoring errors occur when one misconstrues a memory of an imagined event as a memory of a perceived event. These errors may arise as a consequence of similarities between how imagined and perceived events are encoded, as well as similarities among the event features that are reactivated during retrieval (Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981). Such confusions may serve as a basic mechanism for the induction of many types of false memories (Gonsalves & Paller, 2002; Schacter, 1995).

To examine neural events that lead to reality-monitoring errors, we used a behavioral procedure previously shown to produce such errors (Gonsalves & Paller, 2000b), and we monitored brain activity at encoding using whole-brain event-related fMRI with rapid stimulus presentation. We were thus able to investigate the hypothesis that

brain networks involved in generating visual imagery would show greater activity at encoding for imagined events subsequently misremembered as perceived events than for imagined events subsequently remembered accurately as imagined. Our prior electrophysiological results suggested that visual imagery may play a role in the formation of these false memories (Gonsalves & Paller, 2000b), but the data could not support any specific anatomical conclusions. In the present experiment, neuroimaging results implicated brain areas known to play a role in the generation of visual images, thereby providing support for the hypothesis that mental imagery is at the core of the formation of these false memories.

METHOD

Stimuli and Tasks

The stimuli consisted of 350 visually displayed words, 350 color photographs, and 525 spoken words. All words were concrete nouns (mean frequency = 37 occurrences per million, mean concreteness = 556 on a scale from 100 to 700). Spoken words were recorded digitally, and their duration ranged from 240 to 690 ms ($M = 475$ ms, $SE = 4$ ms).

During the study phase (Fig. 1), subjects read object names and were instructed to generate a visual image corresponding to each object. For half of the words, a photographic image of the object was presented 2,000 ms after the word. For the other half, a blank rectangle was presented instead. Subjects were told to make no response to photos and to look at each one while waiting for the next word. Structural MRI scans were acquired after completion of the functional scans of the study phase, which included seven separate runs.

The test phase was conducted outside the scanner and began approximately 20 min after the study phase ended. Subjects heard a randomly ordered sequence of spoken words, 175 of which had been presented visually in the study phase together with a photo, 175 of which had been presented without a photo, and 175 of which had not been presented at all. For each spoken word, subjects decided whether or not they had viewed a photo of the named object during the study phase, and indicated their responses using two keys on a keyboard.

Participants

Eleven volunteers (8 men and 3 women) ages 18 through 28 years ($M = 21$) were recruited from the Northwestern University community. All participants gave informed consent and were screened for MRI contraindications. After the test phase, they were debriefed and paid for their participation.

Imaging Methods

Whole-brain gradient-recalled echo-planar images were collected during the study phase (twenty-four 6-mm slices, TR = 2,000 ms, TE = 40 ms, flip angle = 85°, field of view = 24 cm) using a 1.5-T Siemens Vision scanner. Slices were oriented along the anterior commissure–posterior commissure line (slightly oblique from axial) with a resolution of 3.75 mm × 3.75 mm × 6 mm. In each run, 160 volumes were collected (4 to reach steady state prior to the first stimulus, 150 during the study phase, and 6 to collect hemodynamic responses for the final trials). For anatomical localization, a 3D-FLASH T1-weighted volume was acquired (160 slices, 1-mm axial slices, field of view = 24 cm, 256 × 256 matrix).

Image Analysis

After acquisition, images were co-registered through time using a three-dimensional registration algorithm (Cox, 1996). Functional volumes were spatially smoothed (7.5-mm full-width at half-maximum Gaussian kernel) to improve signal-to-noise ratios and to accommodate anatomical differences across participants. Within each run, voxels were eliminated if the signal changed by more than 10% between two samples or if the mean signal level was below a threshold defined by the inherent noise in the data acquisition. Data from each run were transformed (Collins, Neelin, Peters, & Evans, 1994) to a standard coordinate system (Talairach & Tournoux, 1988) with a final resolution of 2.5 mm^3, and the seven runs were concatenated into a single time series for each participant.

The average response to each trial type was estimated using a general linear model analysis that included indications of the onset of

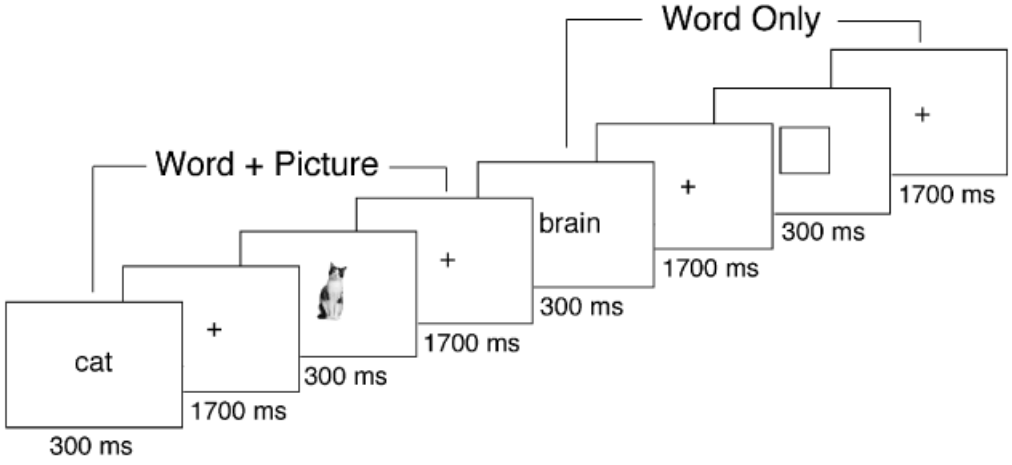

Fig. 1. Procedure used to induce false memories. In the study phase (shown), subjects read names of objects and mentally visualized the referents. Half of the names were followed 2 s later by a photographic representation of the named object. In a surprise memory test given outside the scanner, subjects listened to object names and decided whether they had seen a picture of the object corresponding to each name.

each trial type and several control variables (average signal and linear drift estimated individually for each run, and estimates of corrected motion for each time point to remove signal changes that were correlated with movement). Differences in hemodynamic responses between trial types were estimated by contrasting the average peak response between 4 and 10 s after stimulus onset, accounting for hemodynamic lag. Estimates of differences between trial types were obtained for each participant and combined in a random-effects analysis that identified differences in responses that were consistent across participants. The criterion for identifying a statistically significant activation difference was that a group of voxels in a 500-mm^3 region exhibited a reliable change in activity, $t(10) > 4.0$, $p < .005$ (uncorrected). Monte Carlo simulations using normally distributed noise with 750 time points and 12 participants indicated less than .05 false positives per experiment with this statistical threshold.

RESULTS

Recognition Performance

The two trial types in the study phase were word-plus-photo trials and word-only trials. Endorsement rates in the photo recognition test were 74% ($SE = 3\%$) for word-plus-photo items (i.e., hits), 27% ($SE = 4\%$) for word-only items (i.e., false memories), and 6% ($SE = 2\%$) for new items (i.e., false alarms). False memories occurred significantly more often than false alarms, $t(10) = 7.36$, $p < .001$, $d = 1.75$. Words leading to false memories did not differ in frequency from words leading to correct rejections, but words leading to false memories were on average more concrete, $t(10) = 3.15$, $p < .05$, $d = 1.53$.

Neuroimaging Results

Responses were segregated according to later memory performance to make two contrasts (*false-memory Dm* and *accurate-memory Dm*). False-memory Dm was assessed by responses to word-only trials; responses to words later associated with false memories (i.e., during the test phase, the participant claimed to have seen photos of the named objects) were compared with responses to words later remembered correctly as not having been followed by a photo. Accurate-memory Dm was assessed by responses to word-plus-photo trials; responses to photos subsequently remembered were compared with responses to photos subsequently forgotten. Results are presented in Figure 2 and Table 1.

False-Memory Dm

Three areas showed larger responses to words that were later falsely remembered as having been presented with photos than to words for

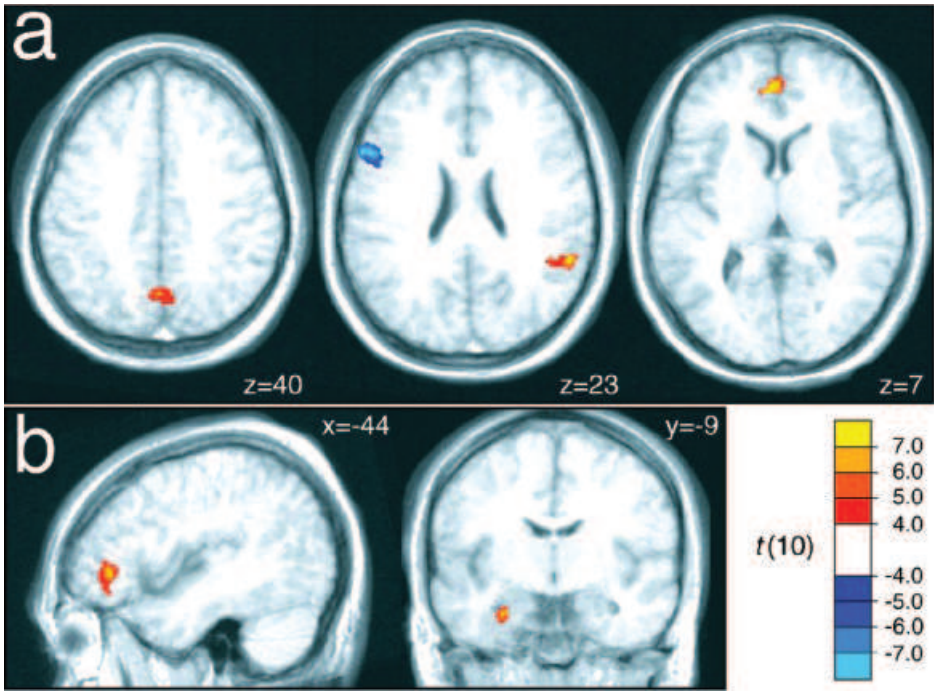

Fig. 2. Neuroimaging results. Three brain areas (a) showed greater responses in the study phase to words that were later falsely remembered as having been presented with photos than to words that were not later misremembered as having been presented with photos. These areas are shown in yellow and orange superimposed on axial images (structural scans averaged across all participants at three levels, with *z*-coordinates listed). From left to right, the images show precuneus, inferior parietal cortex, and anterior cingulate activations, respectively. In addition, a portion of left inferior prefrontal cortex, shown in blue, exhibited greater activity for words that were not later misremembered as having been presented with photos (i.e., correct rejections). In two brain areas (b), responses to photos in the study phase were greater for photos later remembered than for photos later forgotten. These areas, left inferior prefrontal cortex (left image) and left anterior hippocampus (right image), are shown in yellow and orange superimposed on sagittal and coronal images, respectively (structural scans averaged across all participants).

TABLE 1
Clusters Showing Significant Differences in Activation Between Conditions

Brain region	Brodmann Area	Volume (mm^3)	Talairach coordinates		
			x	*y*	*z*
Responses to words					
Larger response for later false memories than for later correct rejections					
Anterior cingulate	32	1,094	−1	48	7
Right inferior parietal cortex	40	875	54	−49	24
Precuneus	7	859	1	−67	40
Larger response for later correct rejections than for later false memories					
Left inferior frontal gyrus	44	1,938	−52	11	26
Responses to photos					
Larger response for later remembered photos than for later forgotten photos					
Left inferior frontal gyrus	47	2,625	−44	36	−5
Left anterior hippocampus	—	656	−28	−9	−27
Larger response for later forgotten photos than for later remembered photos					
Right inferior parietal cortex	40	1,219	43	−49	17
Precuneus	7	969	−12	−49	39

which these errors did not occur: anterior cingulate, precuneus region in the medial parietal lobes, and a right inferior parietal area. A portion of left inferior prefrontal cortex showed larger responses for later correct rejections than for later false memories.

Accurate-Memory Dm

Two areas showed larger responses to photos that were later remembered than to photos that were later forgotten. The left inferior prefrontal activation resembled previously reported subsequent-memory effects for words (Kirchhoff et al., 2000; Wagner, Koutstaal, & Schacter, 1999; Wagner et al., 1998). The left anterior medial temporal activation included the hippocampus. Two areas showed larger responses for forgotten photos relative to remembered photos, precuneus and right inferior parietal cortex.

DISCUSSION

Whereas many studies have examined neural correlates of successful memory encoding (Wagner et al., 1999), our results systematically associate brain activations with the formation of false memories. Our paradigm was designed to create a high degree of overlap between memories for perceived objects and memories for imagined objects. Accordingly, participants occasionally claimed to have seen a photo of an object that they had only imagined, and this happened more than 4 times as often as false alarms for new objects.

In a prior experiment, we performed false-memory and accurate-memory Dm analyses on brain potentials recorded during the study phase (Gonsalves & Paller, 2000b). Accurate-memory Dm effects for remembering photos were widely distributed across the scalp, probably reflecting the activity of multiple brain areas engaged more strongly for photos encoded more effectively. For the false-memory Dm comparison, brain potentials observed 600 to 900 ms after word onset at occipital and parietal scalp locations were predictive of later false remembering; amplitudes were relatively more positive for words that led to later false memories than for words that led to later correct rejections. Similar posterior potentials with maximal amplitudes about 850 ms after word onset were previously associated with generating visual images (Gonsalves & Paller, 2000a). It is therefore tempting to suppose that the false-memory Dm observed in ERP recordings (Gonsalves & Paller, 2000b) reflected particularly vivid visual imagery for some study-phase items, such that the memory representations for those imagined objects, compared with objects not subsequently falsely remembered, more closely resembled memory representations that would have been formed had the corresponding photos been perceived. However, our scalp-recorded ERP data were insufficient for localizing the brain sources of these effects.

With event-related fMRI in the present investigation, we succeeded in observing neural events associated with encoding false memories. Our prediction based on the ERP results was that encoding activity related to the production of such errors would arise from brain areas engaged in the service of visual imagery. Indeed, greater activation of precuneus, right inferior parietal cortex, and anterior cingulate was observed for words leading to later false memories compared with words leading to later correct rejections. These regions have previously been found to activate during visual imagery tasks (Ishai, Ungerleider, & Haxby, 2000; Kosslyn & Thompson, 2000; Suchan et al., 2002), as well as during tasks involving visuospatial working memory or attention (Culham & Kanwisher, 2001; Fletcher et al., 1995; Labar, Gitelman, Parrish, & Mesulam, 1999). This imagery hypothesis is also supported by the fact that words leading to later false memories had significantly higher concreteness ratings than those leading to later correct rejections.

Different patterns of brain activity were predictive of accurate memory for the photos that subjects viewed. In particular, left inferior prefrontal cortex and left hippocampal activations were predictive of subsequent recognition. When similar activations were observed in previous experiments on episodic memory formation, left prefrontal activation was associated with increased phonological and semantic processing for subsequently remembered items, and hippocampal activation with establishing links among multiple cortical regions involved in representing episodes (Paller & Wagner, 2002; Reber et al., 2002; Wagner et al., 1999).

Interestingly, activations in similar right parietal and precuneus regions were identified as predicting false remembering and as predicting picture forgetting. A possible explanation for this correspondence is that both contrasts reflect strong imagery cued by a word. In word-only trials, strong imagery resulted in a false memory, whereas in word-plus-photo trials, it prevented adequate encoding of the picture following the word, leading to poor memory for the picture. Activation in these regions has been previously associated with encoding trials that lead to later forgetting, which may reflect spontaneous imagery that interferes with successful word encoding (Wagner & Davachi, 2001).

Our central conclusions, which pertain to false memories produced during word-only trials in the study phase, are based on the following conjectures. Multiple brain regions are engaged when a subject reads a word, generates a visual image of the corresponding object, and makes a judgment about the typical size of the object. Brain networks for generating and maintaining visual images show an activation level correlated with the strength of visual imagery. Object representations produced when an exceptionally vivid visual image of an object is generated are similar to those produced when that object is actually seen. Accordingly, in the present study, false memories arose when judgments were based on representations that resulted from vivid visual imagery, which occurred in conjunction with precuneus, right parietal, and anterior cingulate activation.

The present results provide a clear demonstration that neural events at encoding can be predictive of later false memories. In the same way, generating a visual image some time between an initial experience and a later memory query may also lead to distorted remembering. Indeed, the intervening time between an important event and when memory is subsequently probed is a likely time for producing memory distortion in many circumstances (Schacter, 1995).

Whereas prior research on the cognitive neuroscience of false memories has emphasized verbal associative paradigms (see reviews by Gonsalves & Paller, 2002; Roediger & McDermott, 2000), our findings add a new dimension to this research. The class of false memories that occur when perceived and imagined events become confused in memory—reality-monitoring errors—is perhaps the most common type of false memory to occur outside the laboratory. A comprehensive understanding of such false memories will thus have wide-ranging ramifications. Our findings show that these errors of memory can be produced in a laboratory setting, and that their likelihood is increased when precuneus, right parietal, and anterior cingulate regions are engaged in the service of visual imagery, producing visual representations that may resemble those that would have been produced if the object had actually been perceived.

Acknowledgments—This research was supported by National Institute of Neurological Disorders and Stroke Grant NS34638 to K.A.P.

REFERENCES

Bartlett, F.C. (1932). *Remembering*. Cambridge, England: Cambridge University Press.

Collins, D., Neelin, P., Peters, T., & Evans, A. (1994). Automatic 3d intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, *18*, 192–205.

Cox, R.W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Culham, J.C., & Kanwisher, N.G. (2001). Neuroimaging of cognitive functions in human parietal cortex. *Current Opinion in Neurobiology*, *11*, 157–163.

Davachi, L., Maril, A., & Wagner, A.D. (2001). When keeping in mind supports later bringing to mind: Neural markers of phonological rehearsal predict subsequent remembering. *Journal of Cognitive Neuroscience*, *13*, 1059–1070.

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and recall in an incidental memory paradigm. *Psychophysiology*, *23*, 298–308.

Fletcher, P.C., Frith, C.D., Baker, S.C., Shallice, T., Frackowiak, R.S., & Dolan, R.J. (1995). The mind's eye—precuneus activation in memory-related imagery. *NeuroImage*, *2*, 195–200.

Gonsalves, B., & Paller, K.A. (2000a). Brain potentials associated with recollective processing of spoken words. *Memory & Cognition*, *28*, 321–330.

Gonsalves, B., & Paller, K.A. (2000b). Neural events that underlie remembering something that never happened. *Nature Neuroscience*, *3*, 1316–1321.

Gonsalves, B., & Paller, K.A. (2002). Mistaken memories: Remembering events that never happened. *The Neuroscientist*, *8*, 391–395.

Henson, R.N., Rugg, M.D., Shallice, T., Josephs, O., & Dolan, R.J. (1999). Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study. *Journal of Neuroscience*, *19*, 3962–3972.

Ishai, A., Ungerleider, L.G., & Haxby, J.V. (2000). Distributed neural systems for the generation of visual images. *Neuron*, *28*, 979–990.

Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Johnson, M.K., & Raye, C.L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Kirchhoff, B.A., Wagner, A.D., Maril, A., & Stern, C.E. (2000). Prefrontal-temporal circuitry for episodic encoding and subsequent memory. *Journal of Neuroscience*, *20*, 6173–6180.

Kosslyn, S.M., & Thompson, W.L. (2000). Shared mechanisms in visual imagery and visual perception: Insights from cognitive neuroscience. In M.S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 975–985). Cambridge, MA: MIT Press.

Labar, K.S., Gitelman, D.R., Parrish, T.B., & Mesulam, M.-M. (1999). Neuroanatomic overlap of working memory and spatial attention networks: A functional MRI comparison within subjects. *NeuroImage*, *10*, 695–704.

Otten, L.J., Henson, R.N., & Rugg, M.D. (2001). Depth of processing effects on neural correlates of memory encoding: Relationship between findings from across- and within-task comparisons. *Brain*, *124*, 399–412.

Paller, K.A., Kutas, M., & Mayes, A.R. (1987). Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography & Clinical Neurophysiology*, *67*, 360–371.

Paller, K.A., & Wagner, A.D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, *6*, 93–102.

Reber, P.J., Siwiec, R.M., Gitelman, D.R., Parrish, T.B., Mesulam, M.-M., & Paller, K.A. (2002). Neural correlates of successful encoding identified using functional magnetic resonance imaging. *Journal of Neuroscience*, *22*, 9541–9548.

Roediger, H.L., & McDermott, K.B. (2000). Distortions of memory. In E. Tulving & F.I.M. Craik (Eds.), *The Oxford handbook of memory* (pp. 149–164). London: Oxford University Press.

Sanquist, T.F., Rohrbaugh, J.W., Syndulko, K., & Lindsley, D.B. (1980). Electrophysiological signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology*, *17*, 568–576.

Schacter, D.L. (1995). *Memory distortion*. Cambridge, MA: Harvard University Press.

Suchan, B., Yaguez, L., Wunderlich, G., Canavan, A.G.M., Herzog, H., Tellmann, L., Homberg, V., & Seitz, R.J. (2002). Neural correlates of visuospatial imagery. *Behavioral Brain Research*, *131*, 163–168.

Talairach, J., & Tournoux, P. (1988). *A coplanar stereotaxic atlas of the human brain*. Stuttgart, Germany: Thieme.

Wagner, A.D., & Davachi, L. (2001). Cognitive neuroscience: Forgetting of things past. *Current Biology*, *11*, R964–R967.

Wagner, A.D., Koutstaal, W., & Schacter, D.L. (1999). When encoding yields remembering: Insights from event-related neuroimaging. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *354*, 1307–1324.

Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R., & Buckner, R.L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*, 1188–1191.

(Received 9/5/03; Revision accepted 2/3/04)