Research Report

Mismaking Memories

Neural Precursors of Memory Illusions in Electrical Brain Activity

Thomas P. Urbach,[1] Sabine S. Windmann,[1] David G. Payne,[2] and Marta Kutas[1]

[1]University of California, San Diego, and [2]Educational Testing Service, Princeton, NJ

ABSTRACT—*Memory illusions—vivid experiences of events that never occurred—could result from inaccuracies either in retrieving memories or in initially storing them. In two experiments, people studied lists of associated words that either did or did not induce later illusory (false) memories of associated but nonpresented lure words. The amplitude of the electrical brain activity during study of words (~500–1,300 ms) that were themselves later correctly remembered reliably distinguished list words that led to such illusory memories from those that did not. This encoding difference associated with subsequent illusory memory (referred to as a DIM)—presumably reflecting item-specific encoding differences—is a neural precursor of memory illusions.*

Memory is fallible, and anyone may have vivid, subjectively compelling memory experiences in which details or even entire events that did not actually occur seem to be remembered—memory illusions (Roediger, 1996). Verbal memory illusions are readily induced in the laboratory when lists of associated words (e.g., *spoke*, *wagon*, *bicycle*, *car*, *turn*, *tire*, *axle*, *round*, *circle*, *roll*) are studied and, in a subsequent memory test, a critical *lure*—a nonpresented semantic associate of the list words (e.g., *wheel*)—is mistakenly "remembered" (Deese, 1959; Roediger & McDermott, 1995).

Encoding processes governing the initial representation and storage of information during memory formation have figured prominently in explanations of memory illusions (e.g., Brainerd & Reyna, 2001; Roediger, McDermott, & Robinson, 1998; Schacter, Normal, & Koutstaal, 1998). Although abundant evidence demonstrates that encoding factors modulate memory-illusion rates (see, e.g., Arndt & Reder, 2003; Cleary & Greene, 2002; Neuschatz, Benoit, & Payne, 2003), studies directly measuring brain activity during encoding are conspicuously absent from the literature. Neuroimaging experiments in which memory illusions are induced using variants of the Deese-Roediger-McDermott (DRM) paradigm just illustrated—even studies investigating encoding influences—have largely focused on brain activity for memory illusions at retrieval (Cabeza, Rao, Wagner, Mayer, & Schacter, 2001; Curran, Schacter, Johnson, & Spinks, 2001; Düzel, Yonelinas, Mangun, Heinze, & Tulving, 1997; Fabiani, Stadler, & Wessels, 2000; Johnson et al., 1997; A.R. Miller, Baratta, Wynveen, & Rosenfeld, 2001; Nessler & Mecklinger, 2003; Nessler, Mecklinger, & Penney, 2001; Schacter, Buckner, Koutstaal, Dale, & Rosen, 1997; Schacter et al., 1996). Like the snapshot of a photo-finish horse race, neurophysiological recordings made as memory illusions occur provide objective measurements of an elusive event but do not reveal how the race was run.

To investigate the relationship between brain activity during encoding and subsequent memory illusions, we conducted two event-related brain potential (ERP) experiments. ERPs—neurally generated potentials elicited by an experimental event and recorded at the scalp—are sensitive to encoding processes, and ERPs recorded while items are studied distinguish those that are subsequently remembered from those that are not (reviewed in Wagner, Koutstaal, & Schacter, 1999). This difference in the ERPs is referred to as a *DM* (difference due to subsequent memory). We extended this kind of subsequent-memory analysis to memory illusions by analyzing the ERPs recorded during the study phase of a DRM paradigm as a function of subsequent memory illusions for the lure. We found that during the study phase, subsequently recognized words that induced a later memory illusion were associated with reduced positive deflections 500 to 1,300 ms poststimulus in comparison with subsequently recognized words that did not induce an illusion (Fig. 1).

Sabine S. Windmann is now at Ruhr-University Bochum, Bochum, Germany. Address correspondence to Thomas P. Urbach, Department of Cognitive Science, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0515; e-mail: turbach@cogsci.ucsd.edu.

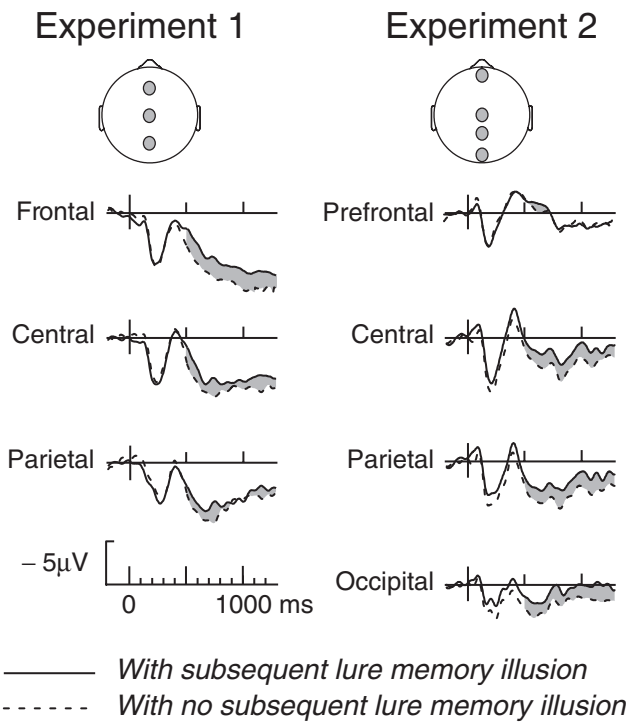

Fig. 1. Event-related potentials elicited by studied words that were correctly recognized in the subsequent recognition memory test, separately for words that did and did not induce false memories that the associated but nonpresented lure words were studied. Waveforms are plotted (negative up) for midline electrodes Fz, Cz, and Pz in Experiment 1 (left) and midline prefrontal, midline central, midline parietal, and midline occipital in Experiment 2 (right).

METHOD

In two experiments, different groups of young, healthy adults were presented with lists of words to study for an upcoming recognition task. Each list consisted of 10 semantic associates of a nonpresented lure. In Experiment 1, each subject studied one of two disjoint sets of 20 DRM lists, and in Experiment 2, each subject studied one of two sets of 40 DRM lists. (Twenty lists drawn from previous experiments conducted by D. Payne and his colleagues were used in both experiments, and the 20 additional lists in Experiment 2 were drawn from Shiffrin, Huber, & Marinelli, 1995, and Stadler, Roediger, & McDermott, 1999.) Words on the recognition memory tests (200 words in Experiment 1 and 400 in Experiment 2) consisted of 4 studied words, 1 nonpresented critical lure, and 5 nonpresented distractors per DRM list. During study, words were presented one at a time on a computer screen. In Experiment 1, 2-min pen-and-paper free-recall tests occurred after every 5th list, with the recognition memory test following all 20 lists. In Experiment 2, after each list was studied, a brief (ca. 45-s) letter-string match-to-sample task was presented, followed by the 10 recognition test words for the immediately preceding list (position of the lure and order of the studied, lure, and distractor words varied). In both experiments, test words were presented one at a time on a computer screen for old/new forced-choice judgment followed by a meta-cognitive judgment. In Experiment 1, a forced-choice *remember* or *know* judgment (Tulving, 1985) followed "old" responses; in Experiment 2, after both "old" and "new" responses, participants made a "sure" response if confident of their answer. Responses on the recognition memory tests were classified as *hits* (studied words correctly judged old), *false alarms* (distractor words incorrectly judged old), and *memory illusions* (lure words incorrectly judged old).

Only electroencephalogram (EEG) recorded during the study phase was analyzed for this report (Experiment 1: 14 scalp locations, details in Neville, Kutas, Chesney, & Schmidt, 1986; Experiment 2: 26 locations, details in Windmann, Urbach, & Kutas, 2002). EEG data were screened for artifacts, with eyeblinks corrected when possible. For each participant separately, ERP waveforms in the experimental conditions were computed by extracting a 2,048-ms epoch of EEG data beginning 500 ms before the onset of each studied word, computing the average across trials at each time point in the epoch, and then digitally low-pass filtering to 15 Hz. The response and ERP data were analyzed only for participants with memory-illusion rates between 10% and 90% and a minimum of 12 trials of artifact-free EEG data per condition: In Experiment 1, 7 of 29 participants (Binghamton University community) were excluded from the analysis of the midline electrode data (1 additional subject was excluded from the analysis of the lateral electrode data because of a recording failure at one of the electrodes); in Experiment 2, 6 of 22 participants (University of California, San Diego, community) were excluded. Both experiments were conducted in accordance with approved guidelines for human-subject research.

RESULTS

In the recognition memory tests, hit rates (mean proportion, with standard errors in parentheses) for the studied words were high: .80 (.02) in Experiment 1 and .91 (.02) in Experiment 2. False alarm rates for the distractor words were low: .10 (.02) in Experiment 1 and .02 (.01) in Experiment 2. As expected, memory-illusion rates for lures were much higher than false alarm rates: .56 (.03) in Experiment 1 and .47 (.06) in Experiment 2. The short (<1 min) study-test retention interval for each list in Experiment 2 is most likely responsible for the higher recognition accuracy and lower memory-illusion rate in this experiment compared with Experiment 1.

Encoding processes were investigated by analyzing study-phase ERPs elicited by words that were correctly recognized in the subsequent memory test. ERPs for words that did not lead to subsequent memory illusions were systematically more positive than ERPs for words that did (Fig. 1); we refer to this ERP effect as a *DIM* (difference in subsequent illusory memory). The DIM began at approximately 400 ms in Experiment 1 and even

earlier in Experiment 2, with somewhat different scalp distributions in the two experiments.

For statistical analysis, mean potentials at midline electrodes were measured relative to a mean amplitude in the 200-ms interval immediately preceding stimulus onset. These measures were taken in four successive time windows: 100–300 ms (P2), 300–500 ms (N4), 500–800 ms (late positive complex, or LPC), and 800–1,300 ms (slow wave, or SW). Repeated measures analyses of variance ($\alpha = .05$ on Huynh-Feldt ε-adjusted *df*) were conducted with two levels of word type and three levels of frontal plane (frontal, central, parietal) in Experiment 1, four levels of frontal plane (prefrontal, central, parietal, occipital) in Experiment 2. Effect size (η_p^2) was calculated as $SS_{effect}/(SS_{effect} + SS_{error})$. Because different DRM lists induce memory illusions at different rates (Stadler et al., 1999), the individual words from different lists occurred in different proportions in the two conditions used to compute the within-subjects DIM effect. To address this potential confound, we computed within-word DIM ERP effects by subtracting (a) the average ERP elicited by each word when it was studied as part of a DRM list that induced a later memory illusion from (b) the average ERP elicited by the same word when it was studied as part of a list that did not induce a later memory illusion. By averaging ERPs for words with two or more trials of each type, the within-word DIM effect (Fig. 2) could be analyzed for 104 (65%) of the 160 study-phase words in Experiment 1 that appeared in the later recognition test and 190 (59%) of the 320 such words in Experiment 2.

Analysis of the within-subjects DIM at midline electrodes found no statistically significant effects in the P2 or N4 windows. In the LPC and SW windows in both experiments, ERPs for subsequent hits that did not lead to memory illusions were reliably more positive than ERPs for subsequent hits that did. In Experiment 1, the effect was broadly distributed over the midline electrodes; the trend toward greater relative positivity at the frontal site in comparison with the central and parietal sites was not statistically reliable, $F(2, 42) < 1$. Collapsing across the three midline electrodes, the amplitude of the effect was 1.60 μV in the LPC window, $F(1, 21) = 8.16$, $p = .01$, $\eta_p^2 = .28$, and 1.27 μV in the SW window, $F(1, 21) = 4.78$, $p = .04$, $\eta_p^2 = .19$. The corresponding within-word DIM main effect was in the expected positive direction, 1.05 μV in the LPC window, $t(103) = 1.46$, $p = .075$ (one-tailed), $\eta_p^2 = .02$, and 1.13 μV in the SW window, $t(103) = 1.50$, $p = .069$ (one-tailed), $\eta_p^2 = .02$.

In Experiment 2, ERPs to subsequent hits that did not lead to later memory illusions were again more positive in both the LPC and SW windows than were ERPs to subsequent hits that did not

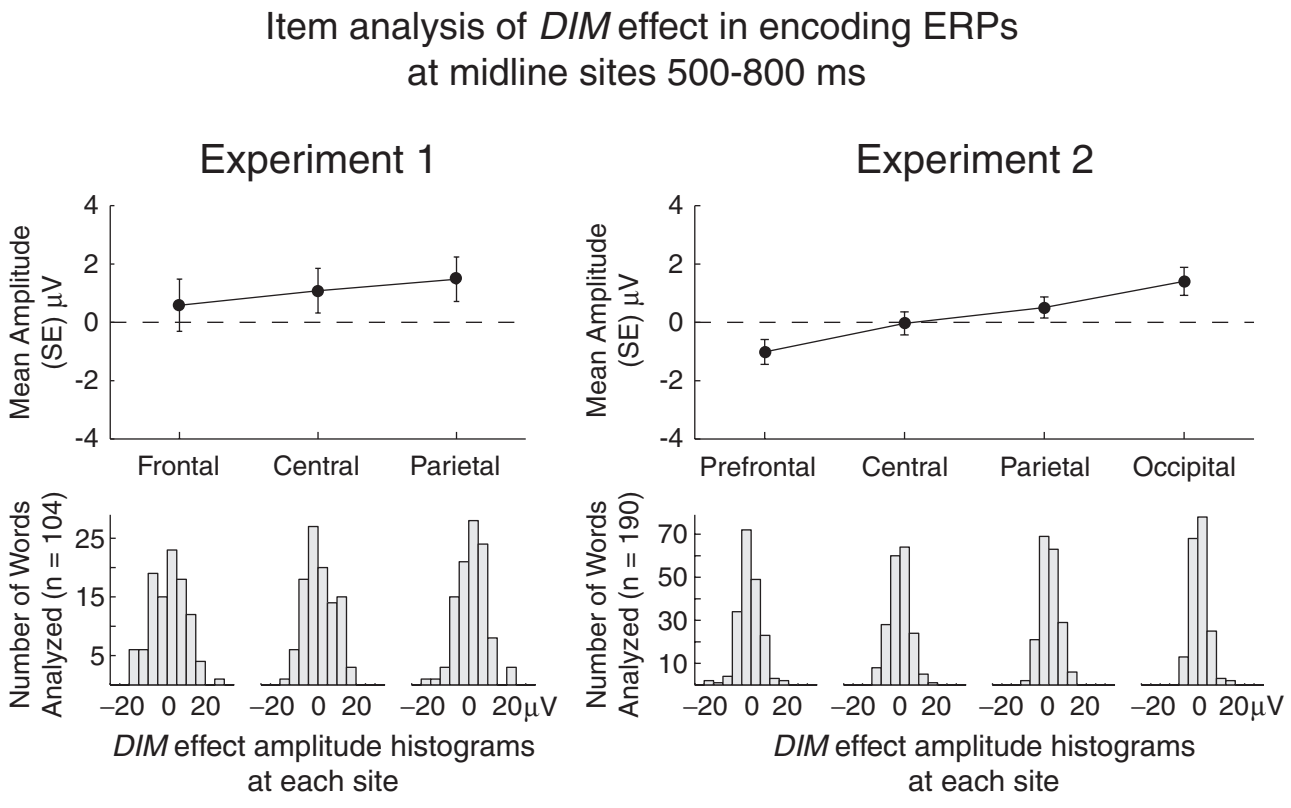

Fig. 2. Results of the item analysis of the difference in amplitude of the event-related potentials (ERPs) elicited by individual items when their lists did and did not induce a later memory illusion (referred to as the DIM effect). The top panels show the grand mean amplitude (with standard error) of these individual-word DIM effects at each electrode site in Experiments 1 and 2. The bottom panels present the corresponding histograms for the 104 words (65%) in Experiment 1 and the 190 words (59%) in Experiment 2 that were analyzed.

lead to memory illusions. The effect was broadly distributed over the midline electrodes; although the effect was larger posteriorly than anteriorly, the interaction effect for word type and electrode location was not significant in either window, $F(3, 45) < 1.4$. Collapsing over the four midline electrodes, the DIM amplitude was 0.98 μV in the LPC window, $F(1, 15) = 5.88$, $p = .028$, $\eta_p^2 = .28$, and 0.79 μV in the SW window, $F(1, 15) = 5.32$, $p = .036$, $\eta_p^2 = .26$. The corresponding within-word DIM main effect was 0.21 μV in the LPC window, $t(189) = 0.76$, n.s., and 0.48 μV in the expected positive direction in the SW window, $t(189) = 1.72$, $p = .04$ (one-tailed), $\eta_p^2 = .02$.

In Experiment 2, the within-word DIM was positive at the posterior electrodes and negative at the prefrontal electrode. In the LPC window, this unexpected −1.02 μV effect at the prefrontal electrode was reliable, $t(189) = -2.386$, $SD = 5.87$, $p = .018$ (two-tailed, unadjusted). This prefrontal negativity may be an idiosyncratic property of the subset of words analyzed, because it was not evident in the within-subjects analysis of the DIM effect using the entire stimulus set. With this exception in Experiment 2, the within-word DIM effects at the midline electrodes 500 to 1,300 ms poststimulus accord well with the DIM effects calculated within subjects over the entire stimulus set. The loss of statistical power in the within-word analysis of the DIM effect, reflected in the low t values and negligible amount of variability explained, is not surprising because computing individual-word ERP averages over small numbers of between-subjects trials results in high interitem variability.

The results from the lateral electrode analyses did not differ materially from the results from the midline analyses. In Experiment 1, there were no lateral asymmetries, and the only reliable effect was an interaction between word type and electrode location in the frontal plane in the P2 window. The DIM effect at lateral electrodes was 0.23 μV at the frontal scalp sites (F7, F8), 0.10 μV at the anterior temporal sites (ATL, ATR), −0.32 μV at temporal sites (TL, TR), −0.37 μV at sites over Wernicke's area and its right-hemisphere homologue (WR, WL), and −0.65 μV at parietal scalp sites (P3, P4), $F(4, 80) = 4.013$, $p = .011$, $\eta_p^2 = .17$. This P2 effect was not found in Experiment 2. In Experiment 2, the medial centro-parietal maximum of the positivity resulted in significant interactions between word type and electrode locations for the LPC and SW windows. The scalp distributions of potentials in Experiment 2 are illustrated in Figure 3. The DIM amplitude 500 to 1,300 ms poststimulus was maximal at parietal scalp sites (Fig. 3c), although η_p^2 calculated at each electrode was greatest at left fronto-central electrodes, where more than 40% of the variance was explained (Fig. 3d).

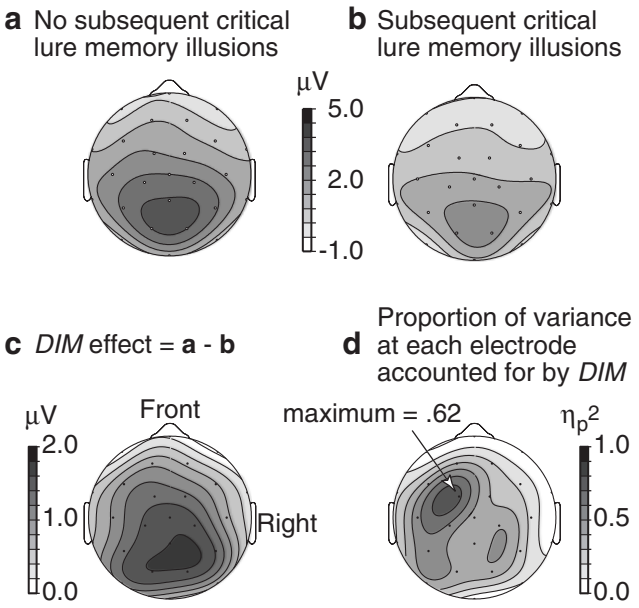

Fig. 3. Results from recordings at 28 electrodes 500 through 1,300 ms after word onset in the study phase of Experiment 2: scalp distribution of (a) potentials for subsequently recognized studied words that did not lead to a subsequent memory illusion; (b) potentials for subsequently recognized studied words that did lead to a subsequent memory illusion; (c) the DIM effect (difference associated with subsequent memory), calculated as potentials in (a) minus potentials in (b); and (d) proportion of variance in the encoding potentials for subsequently recognized words accounted for by DIM as given by the value of η_p^2 calculated at each electrode. Results are represented as spline interpolated contour maps projected on a hemisphere.

DISCUSSION

We found that even with encoding task demands held constant, at least some of the processing related to subsequent memory illusions occurred (or failed to occur) 500 to 1,300 ms after the experience began to be neurally represented. In contrast with measures of retrieval performance following encoding manipulations, which allow indirect inferences about encoding processes, the DIM is a neural precursor of memory illusions that is observed while the encoding processes themselves unfold. Though the functional significance of the DIM is presently uncertain, it may reflect the extent to which item-specific information is encoded. Previous ERP memory research has found that encoding strategies or tasks that lead to better subsequent recognition memory performance tend to be associated with more positive DM effects at encoding and that the specific scalp distributions of these effects vary as a function of encoding and retrieval task demands (Wagner et al., 1999).

Converging evidence comes from a remember/know recognition memory experiment (Friedman & Trott, 2000) that found encoding ERPs after about 500 ms poststimulus were more positive for subsequently remembered words than for subsequently known words. Although subsequently remembered versus subsequently known words do not always exhibit a DM positivity (Mangels, Picton, & Craik, 2001, focused attention condition; Smith, 1993), to the extent that remember judgments are supported by retrieval of episode-specific information, Friedman and Trott's finding suggests an association between

relative positivity in the DM and encoding specific aspects of the stimuli. The DIM positivity, too, may reflect encoding specific features of the studied words; such encoding, in turn, improves the ability to discriminate between actually studied words and related but nonpresented critical lures. In a different memory-distortion paradigm, Gonsalves and Paller (2000) found that ERPs to studied words presented without pictures were more positive over visual cortex if the words were misremembered as having been presented along with a picture. This apparently discrepant result is consistent and complementary if the positivity reflects encoding of item-specific information: Robust encoding of item-specific visual images could tend to increase source-confusion errors by making the experience of the (imaged) word less discriminable from the experience of an actual picture. In the DRM paradigm, encoding item-specific information about the studied words could make the stored representations of the individual words more distinct and enable participants to better discriminate studied words from lures.

The suggestion that the DIM is associated with encoding of item-specific information is a working hypothesis that requires further investigation, but the DIM has important implications independent of this interpretation. The issue of how encoding and retrieval processes conspire to produce verbal memory illusions has attracted considerable theoretical discussion. Whereas it is widely accepted that encoding processes play a prominent role in verbal false memories (Roediger et al., 1998; but see M.B. Miller & Wolford, 1999), the DIM effect provides direct evidence for this received view based on measurements of the actively encoding brain. The DIM effect—a snapshot taken as the mnemonic horses break from the gate—confirms that encoding processes play a role in the etiology of verbal memory illusions and offers an intriguing initial glimpse into their neural time course.

Acknowledgments—Experiment 1 was conducted while T.P.U. and D.G.P. were at State University of New York, Binghamton. This research was supported in part by a postdoctoral scholarship of the German academic exchange service (DAAD; Bonn, Germany) to S.S.W. and by National Institute of Child Health and Human Development Grant HD22614 and National Institute on Aging Grant AG08313 to M.K. For their assistance, we thank J. Blackwell, J. Cagle, and P. Krewski.

REFERENCES

Arndt, J., & Reder, L.M. (2003). The effect of distinctive visual information on false recognition. *Journal of Memory and Language*, *48*, 1–15.

Brainerd, C.J., & Reyna, V.F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior*, *28*, 41–100.

Cabeza, R., Rao, S.M., Wagner, A.D., Mayer, A.R., & Schacter, D.L. (2001). Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proceedings of the National Academy of Sciences, USA*, *98*, 4805–4810.

Cleary, A.M., & Greene, R.L. (2002). Paradoxical effects of presentation modality on false memory. *Memory*, *10*, 55–61.

Curran, T., Schacter, D.L., Johnson, M.K., & Spinks, R. (2001). Brain potentials reflect behavioral differences in true and false recognition. *Journal of Cognitive Neuroscience*, *13*, 201–216.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Düzel, E., Yonelinas, A.P., Mangun, G.R., Heinze, H.J., & Tulving, E. (1997). Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences, USA*, *94*, 5973–5978.

Fabiani, M., Stadler, M.A., & Wessels, P.M. (2000). True but not false memories produce a sensory signature in human lateralized brain potentials. *Journal of Cognitive Neuroscience*, *12*, 941–949.

Friedman, D., & Trott, C. (2000). An event-related potential study of encoding in young and older adults. *Neuropsychologia*, *38*, 542–557.

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

Johnson, M.K., Nolde, S.F., Mather, M., Kounios, J., Schacter, D.L., & Curran, T. (1997). The similarity of brain activity associated with true or false recognition memory depends on test format. *Psychological Science*, *8*, 250–257.

Mangels, J.A., Picton, T.W., & Craik, F.I.M. (2001). Attention and successful episodic encoding: An event-related potential study. *Cognitive Brain Research*, *11*, 77–95.

Miller, A.R., Baratta, C., Wynveen, C., & Rosenfeld, J.P. (2001). P300 latency, but not amplitude or topography, distinguishes between true and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 354–361.

Miller, M.B., & Wolford, G.L. (1999). The role of criterion shift in false memory. *Psychological Review*, *106*, 398–405.

Nessler, D., & Mecklinger, A. (2003). ERP correlates of true and false recognition after different retention delays: Stimulus- and response-related processes. *Psychophysiology*, *40*, 146–159.

Nessler, D., Mecklinger, A., & Penney, T.B. (2001). Event related brain potentials and illusory memories: The effects of differential encoding. *Cognitive Brain Research*, *10*, 283–301.

Neuschatz, J.S., Benoit, G.E., & Payne, D.G. (2003). Effective warnings in the Deese-Roediger-McDermott false-memory paradigm: The role of identifiability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 35–41.

Neville, H.J., Kutas, M., Chesney, G., & Schmidt, A.L. (1986). Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. *Journal of Memory and Language*, *25*, 75–92.

Roediger, H.L., III. (1996). Memory illusions. *Journal of Memory and Language*, *35*, 76–100.

Roediger, H.L., III, & McDermott, K.B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814.

Roediger, H.L., III, McDermott, K.B., & Robinson, K.J. (1998). The role of associative processes in creating false memories. In M.A. Conway, S.E. Gathercole, & C. Cornoldi (Eds.), *Theories of memory II* (pp. 187–246). Hove, England: Psychological Press.

Schacter, D.L., Buckner, R.L., Koutstaal, W., Dale, A.M., & Rosen, B.R. (1997). Late onset of anterior prefrontal activity during true

and false recognition: An event-related MRI study. *NeuroImage*, *6*, 259–269.

Schacter, D.L., Normal, K.A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 289–318.

Schacter, D.L., Reiman, E., Curran, T., Yun, L.S., Bandy, D., McDermott, K.B., & Roediger, H.L. (1996). Neuroanatomical correlates of veridical and illusory recognition memory: Evidence from positron emission tomography. *Neuron*, *17*, 267–274.

Shiffrin, R.M., Huber, D.E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 267–287.

Smith, M.E. (1993). Neurophysiological manifestations of recollective experience during recognition memory judgments. *Journal of Cognitive Neuroscience*, *5*, 1–13.

Stadler, M.A., Roediger, H.L., III, & McDermott, K.B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *27*, 494–500.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology-Psychologie Canadienne*, *26*, 1–12.

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 Series B: Biological Sciences*, *354*, 1307–1324.

Windmann, S.S., Urbach, T.P., & Kutas, M. (2002). Cognitive and neural mechanisms of decision biases in recognition memory. *Cerebral Cortex*, *12*, 808–817.

(Received 8/8/03; Revision accepted 12/11/03)