# ERPS PREDICTIVE OF SUBSEQUENT RECALL AND RECOGNITION PERFORMANCE *

Ken A. PALLER **, Gregory McCARTHY and Charles C. WOOD

*Neuropsychology Laboratory, Veterans Administration Medical Center, West Haven, CT 06516, U.S.A. and Departments of Neurology and Psychology, Yale University, New Haven, CT 06520, U.S.A.*

By exploiting measures of information processing complementary to those obtained from behavioral studies, electrophysiological studies of human memory may provide insights into the cognitive processes associated with encoding. In the present experiment, subjects viewed words under incidental learning conditions in which each word required a two-choice decision based on semantic criteria (*interesting/uninteresting* or *edible/inedible*). Memory for those words was subsequently assessed by a free recall test and then a recognition test. Event-related brain potentials elicited in response to the original presentation of each word were found to differ as a function of later memory performance. Over the 400–800 ms latency range, responses to remembered words were positive relative to responses to forgotten words, especially for recall. These electrophysiological differences are interpreted as reflections of processes that correlated with encoding.

## 1. Introduction

Previous studies have shown that event-related potentials (ERPs) elicited by words can be predictive of subsequent memory performance for those words (Fabiani, Karis, & Donchin, 1985, 1986; Friedman & Sutton, 1987; Karis, Fabiani, & Donchin, 1984; Neville, Kutas, Chesney, & Schmidt, 1986; Paller, Kutas, & Mayes, 1987; Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980). For example, Paller et al. (1987) used an incidental learning paradigm in which subjects made two-choice semantic and nonsemantic judgments about words and later were given recall and recognition tests. ERPs elicited by each word during acquisition were sorted according to whether that word was or was not remembered. ERPs to remembered words differed significantly from ERPs to words that were not remembered, in that the former were positive relative to the latter over the latency range 400–800 ms following word onset.

* Supported by the Veterans Administration and NIMH Grant MH-05286. We thank Joe Jasiorkowski, Liz Roessler, and Mary Pearson for technical assistance.

** Address for correspondence: Ken Paller, Neuropsychology Laboratory, 116B1, VA Medical Center, West Haven, CT 06516, U.S.A.

To avoid prejudging the relationship between this ERP difference and ERP components such as P300, Paller et al. (1987) introduced the term *Dm* to refer to ERP *D*ifferences based on subsequent *m*emory performance.

This type of ERP difference has been investigated using several different experimental paradigms. Neville et al. (1986) used an incidental learning paradigm in which subjects decided whether each word was congruous or incongruous with a preceding phrase. Results from a recognition test were used to calculate Dm, which began 250 ms after the onset of congruous words and 550 ms after the onset of incongruous words. Sanquist et al. (1980) required subjects to decide whether each pair of words was the same or different based on orthographic, phonemic, or semantic criteria. ERPs were found to differ as a function of later recognition performance, although only a small number of trials were available for this comparison. Fabiani et al. (1986) reported ERP differences related to later recall performance using an incidental learning paradigm in which a discrimination between male and female names was required. These effects were recorded at Fz, Cz, and Pz and were largest for three subjects not using complex mnemonic strategies. For the other nine subjects (who did use complex mnemonic strategies), Dm was apparent only at Fz. This pattern of results has been substantiated in a follow-up study in which mnemonic strategies were manipulated (Fabiani et al., 1985). Friedman and Sutton (1987) used a paradigm in which pictures of common objects were presented and subjects decided whether each picture had or had not been presented previously. Once again, ERPs in this study differed as a function of later recognition performance. Comparing these seven experiments is difficult because different ERP analysis techniques, electrode locations, and task requirements were employed. Nevertheless, each of the seven experiments found that ERPs to remembered words were more positive than ERPs to words that were forgotten.

Dm may be of value in the study of memory since it is derived from recordings made during the acquisition stage and may reflect encoding or related memory processes occurring at that time. Furthermore, knowledge of the neural systems that generate Dm may provide information about the neural substrates of memory. However, the utility of Dm remains in question because some analyses of ERPs averaged as a function of subsequent memory performance found no Dm. In a study by Johnson, Pfefferbaum, and Kopell (1985), subjects were instructed to memorize 75-word lists in a repeated study-test procedure. Peak latencies of latency-adjusted waveforms predicted later recognition performance, but amplitude measures did not differ significantly. Paller, Kutas, Shimamura, and Squire (1987a, 1987b) used an incidental learning paradigm with a four-choice concreteness judgment task. Dm was apparent for stem-completion priming (a type of memory not requiring explicit retrieval) but not for recognition in one experiment nor for free recall or cued recall in the other experiment. Such negative findings may reflect an inherent unreliability of Dm or they may reflect systematic differences in the conditions of memory acquisition or testing in these experiments. A better understanding of the necessary and sufficient conditions for the elicitation of Dm is therefore important for Dm to be of value. In this study, we attempted to replicate the findings of Paller et al. (1987), including comparisons between recall and recognition. The incidental learning paradigm was structured so that the same word lists could be used in subsequent studies comparing recognition to stem-completion priming (Paller, Wood, & McCarthy, 1988).

## 2. Method

Subjects were 10 right-handed adults between 19 and 35 years old. Words were presented on a video monitor for 200 ms at a constant rate of one word every 2 s. Ten lists, each consisting of 24 concrete nouns, were presented in the same order to each subject. Subjects were not told that memory tests would be given. Instead, five subjects were assigned task I for even-numbered lists and task E for odd-numbered lists; the other five subjects were given the opposite assignments. Task I required each word to be judged either *interesting* or *uninteresting*. Task E required each word to be judged either *edible* or *inedible*. Subjects responded by pressing buttons with their right hand. For task I, subjects were told to distribute their responses roughly equally between the two response categories. For most analyses, data were combined across the two tasks. Data for the two words at the beginning and end of each list were excluded to minimize primacy and recency effects. A distraction task (counting backwards by threes for 1 min) was assigned after the tenth list to minimize the influence of immediate memory on the two memory tests, though neither test was expected.

After the distraction task, a free recall test was given. Subjects were allotted 15 min in which to write down words from the preceding 10 lists. After a further 30–60 min delay during which subjects performed tasks requiring tone-frequency discriminations, a yes-no recognition test was given. The recognition test was a randomly ordered list of 200 previously presented words and 800 new words. Subjects were instructed to circle each word that they believed had been presented earlier and were informed that a monetary bonus would be awarded for good performance on this test. Subjects were instructed not to go back over their answers. The mean time to complete the recognition test was 35 min.

EEG was recorded from silver disk electrodes at the mastoids and 10 scalp locations (Fz, Cz, Pz, Oz, T3, T4, T5, T6, P3, P4), with a common sterno-vertebral reference. To eliminate residual EKG contamination, the reference was recalculated off-line to the average of left mastoid and right mastoid recordings. The bandpass was 0.1–100 Hz (−3 dB) and data were sampled every 6 ms. ERPs during acquisition were averaged beginning 100 ms prior to word onset and continuing for 1436 ms. Seven percent of the trials were excluded due to contamination by horizontal or vertical EOG artifact.

## 3. Results

Recall scores ranged from 7% to 20% correct, with a mean of 12%. ERPs elicited by words later recalled were more positive than ERPs elicited by words later not recalled, as shown in fig. 1. The average amplitude was measured at all 10 electrodes over the 400–800 ms latency range (as in the study of Paller et al., 1987). This measurement differed significantly as a function of later recall, Recall main effect, $F(1,9) = 10.6$, $p < 0.01$. In addition, the Recall × Electrode interaction was marginally significant, $F(9,81) = 2.7$, $p < 0.078$ (using the Geisser-Greenhouse correction), reflecting smaller amplitudes for Dm at the four temporal sites (average of 0.6 $\mu$V) compared to those at the other sites (average of 1.9 $\mu$V). The distribution of these ERP differences along the midline, nearly equal at Fz, Cz, and Pz, did not correspond to the parietal-maximum distribution usually ascribed to P300.

To investigate the time-course of these effects, separate analyses were conducted over consecutive 100-ms intervals. The Recall main effect was significant for three intervals: 400–500 ms poststimulus, $F(1,9) = 14.3$, $p <$

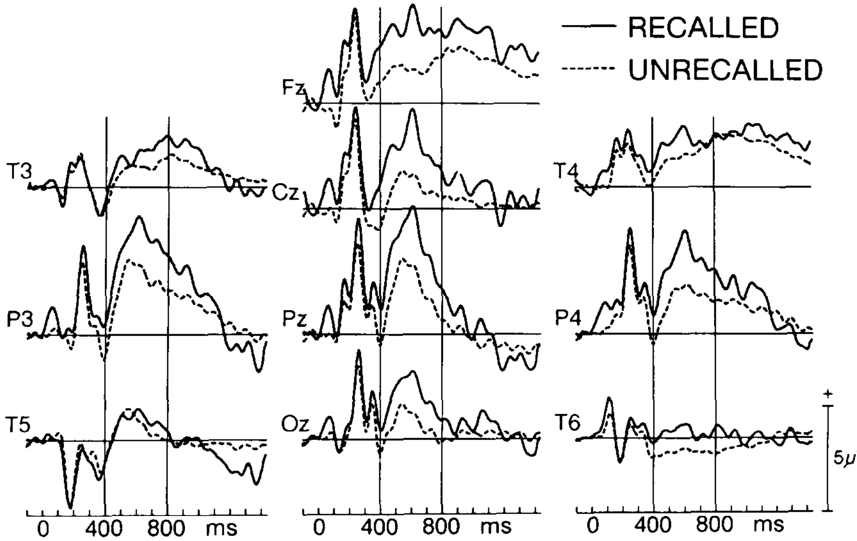

Fig. 1. ERPs averaged on the basis of subsequent performance on the free recall test.

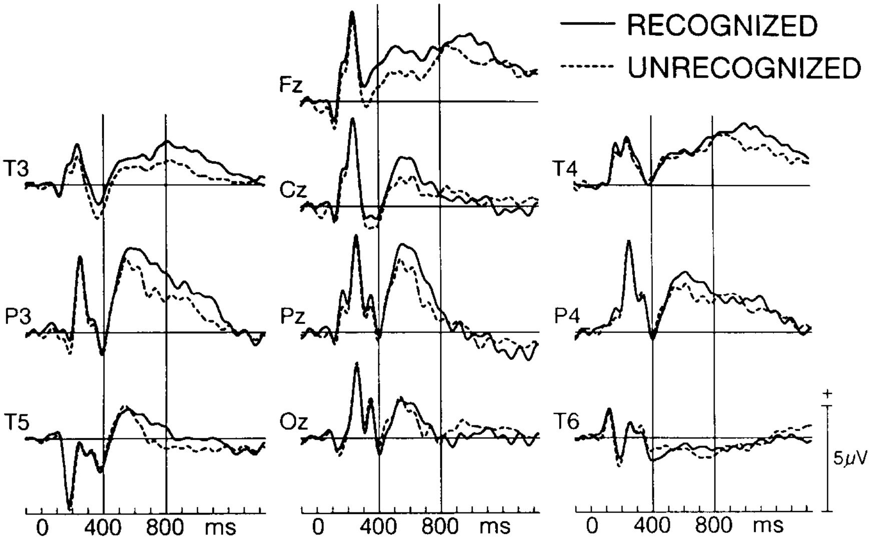

Fig. 2. ERPs averaged on the basis of subsequent performance on the recognition test.

0.005; 500–600 ms poststimulus, $F(1,9) = 9.0$, $p < 0.016$; 600–700 ms poststimulus, $F(1,9) = 13.1$, $p < 0.006$; and nonsignificant for all other intervals.

Recognition scores ranged from 33% to 71% correct, with a mean of 53%. A measure of recognition sensitivity, $d'$, ranged from 0.9 to 3.4 and averaged 1.7. As shown in fig. 2, ERPs appeared to differ as a function of later recognition performance. Measurements of these differences over the 400–800 ms latency range failed to reach statistical significance, although marginal effects were found if measurements were restricted to other portions of the epoch (e.g., the 600–800 ms latency range, $F(1,9) = 3.8$, $p < 0.084$) or to the Fz electrode location, $F(1,9) = 7.3$, $p < 0.025$.

In general, the amplitude of Dm for recall was greater than the amplitude of Dm for recognition. Mean amplitude measurements over the 400–800 ms latency range were analyzed from both tests combined in a three-way ANOVA (Memory × Test × Electrode). Across tests, ERP measurements differed significantly as a function of later memory performance, $F(1,9) = 8.6$, $p < 0.017$. Differences between the tests, however, contributed to a marginal effect for the Memory × Test interaction, $F(1,9) = 4.8$, $p < 0.057$. At midline electrodes, for example, measurements of Dm averaged 2.0 μV for recall and 0.6 μV for recognition.

Results were also analyzed as a function of task performance. Both recall and recognition performance were better for affirmative decisions than for negative decisions. For task I, words rated *interesting* were remembered better than words rated *uninteresting* (19% vs. 11% for recall, $F(1,9) = 9.5$, $p < 0.014$;

65% vs. 56% for recognition, $F(1,9) = 6.2$, $p < 0.035$). For task E, words rated *edible* were remembered better than words rated *inedible* (23% vs. 8% for recall, $F(1,9) = 23.0$, $p < 0.001$; 74% vs. 43% for recognition, $F(1,9) = 60.1$, $p < 0.001$). Parallel effects of task decision were apparent in the ERPs. The measurement over the 400–800 ms interval was larger for affirmative decisions than for negative decisions both in task I (1.9 $\mu$V vs. 1.0 $\mu$V) and in task E (5.1 $\mu$V vs. 1.5 $\mu$V), though the difference was significant only for task E, $F(1,9) = 8.9$, $p < 0.016$. The larger ERP effects in task E may have been related to the fact that only 16% of the words were rated *edible* in task E, whereas 44% of the words were rated *interesting* in task I.

## 4. Discussion

As in previous studies, ERPs were predictive of later memory performance in that ERPs recorded during acquisition differed depending on whether words were remembered or forgotten. Words subsequently recalled elicited ERPs that were more positive (especially over the 400–700 ms latency range) compared to ERPs elicited by words not recalled. Similar ERP differences were evident between recognized and unrecognized words, although Dm for recognition reached only marginal statistical significance in the present group of 10 subjects. As noted above, Dm for recognition was about 1/3 the size of Dm for recall; across-subject variability of Dm for recognition was about 1/2 that of Dm for recall (mean square errors were 5.4 and 9.2, respectively). Thus, the difference in statistical significance between the two tests appears a consequence not of greater variability but of smaller size. Despite its smaller size, we regard Dm for recognition as a reliable effect, based on evidence for this type of effect in several experiments (e.g., Friedman & Sutton, 1987; Neville et al., 1986; Paller et al., 1987; Sanquist et al., 1980), including two recent experiments using nearly the same design as in the present experiment (Paller et al., 1988). In addition, affirmative decisions in both task E and task I were associated with better memory and greater positivity than were negative decisions, confirming and extending the results reported by Paller et al. (1987).

Because Dm was measured during initial word presentation, it is presumed to reflect encoding or associated processes that occurred at that time. These processes could be specific to the representation of each word or they could be nonspecific processes (akin to arousal) that were correlated with encoding. However, since Dm is defined by sorting trials based on subsequent memory performance, its magnitude also may have been influenced by rehearsal, interference, retrieval, or other processes that occurred subsequent to the time Dm was recorded. This possibility could account for the smaller Dm with the recognition test than with the recall test if factors such as guessing played a greater role in recognition than in recall, as is likely. In other words, recall performance may have provided a criterion for sorting trials that was more sensitive to differences in encoding effectiveness across words (assuming that Dm reflected some of this effectiveness, which could correspond to encoding strength). In the extreme, it is possible that recognition performance did not provide any sensitivity as a sorting criterion beyond that provided by recall performance, but instead diluted the sensitivity. Dm for recognition was in fact very small when all recalled words were excluded. However, previous results showed that recognition confidence ratings enhanced ERP differences associated with later recognition performance (Paller et al., 1987). Thus, it is likely that both Dm for recall and Dm for recognition depended on encoding effectiveness or strength, rather than on some process unique to recalled words. We assume that each word develops a particular encoding strength during learning. By this reasoning, Dm would be proportional to the difference in the encoding strength between the categories of remembered and forgotten words, which in turn depends on processing requirements at acquisition as well as on the variability in the measure and the sensitivity of the memory test.

An alternative way to explain the finding that Dm for recall was larger than Dm for recognition is that some subjects showed no Dm for recognition because their recognition performance was poor. However, the amplitude of Dm did not correlate with memory performance across subjects for either recall (number of words recalled, $r = 0.12$) or recognition ($d'$ for recognition, $r = 0.12$). It appears that differences in Dm between recall and recognition could not have been due to a recognition floor effect across subjects.

Another way to account for the larger Dm for recall than for recognition is to suppose that Dm reflected an encoding process resembling the process Mandler (1980) has termed “elaboration.” In this scenario, explicit retrieval is made possible by relationships that are established by elaboration between the to-be-remembered word and other information. Whereas recall and recognition performance are supported by retrieval, the familiarity judgements in recognition tests can also be supported by other processes (“activation” or “integration” in Mandler’s terminology, “perceptual fluency” according to Jacoby & Dallas, 1981). The cues given in recognition tests aid retrieval and make recognition less dependent on elaborative processing at acquisition. The finding that Dm was larger when semantic processing was required at acquisition (Paller et al., 1987) is consistent with the idea that Dm reflects elaboration. However, other studies have shown that Dm at Pz is diminished in subjects engaging in certain types of complex semantic processing (Fabiani et al., 1985; Karis et al., 1984). One way to reconcile these results is to suppose that these types of complex mnemonic strategies are like rehearsal episodes that minimize Dm because they occur some time after the ERP is recorded, whereas elaboration will contribute to Dm if it occurs soon enough after the presentation of the word to be reflected in the ERP associated with that word.

To summarize, ERPs elicited several hundred ms after word onset differed as a function of later memory performance. Dm for recall and Dm for recognition were largest near the midline and approximately equal at Fz, Cz, and Pz. Consonant with most previous reports of this type (Fabiani et al., 1986; Friedman & Sutton, 1987; Karis et al., 1984; Neville et al., 1986; Paller et al., 1987; Sanquist et al., 1980), ERPs to words later remembered were more positive than ERPs to words not remembered. Continuing to characterize the full range of conditions under which Dm is and is not elicited should sharpen our understanding of the processes underlying Dm and of the influence of these processes on memory performance.

## References

Fabiani, M., Karis, D., & Donchin, E. (1985). Effects of mnemonic strategy manipulation in a von Restorff paradigm. *Psychophysiology, 22,* 588–589, Abstract.

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

Friedman, D., & Sutton, S. (1987). Event-related potentials during continuous recognition memory. In R. Johnson, Jr., J.W. Rohrbaugh, & R. Parasuraman (Eds.), *Current trends in event-related potential research. Electroencephalography and Clinical Neurophysiology,* (Suppl. 40), 316–321. Amsterdam: Elsevier.

Jacoby, L.L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

Johnson, R., Jr., Pfefferbaum, A., & Kopell, B.S. (1985). P300 and long-term memory: Latency predicts recognition performance. *Psychophysiology, 22,* 497–507.

Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the von Restorff effect. *Cognitive Psychology, 16,* 177–216.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Neville, H., 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.

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

Paller, K.A., Kutas, M., Shimamura, A.P., & Squire, L.R. (1987a). Brain responses to concrete and abstract words reflect processes that correlate with later performance on a test of stem-completion priming. In R. Johnson, Jr., J.W. Rohrbaugh, & R. Parasuraman (Eds.), *Current trends in event-related potential research. Electroencephalography and Clinical Neurophysiology,* (Suppl. 40), 360–365. Amsterdam: Elsevier.

Paller, K.A., Kutas, M., Shimamura, A.P., & Squire, L.R. (1987b, June). *ERPs predictive of later performance on a stem-completion priming test.* Paper presented at the Fourth International Conference on Cognitive Neuroscience, Paris-Dourdan, France.

Paller, K.A., Wood, C.C., & McCarthy, G. (1988). Brain potentials predictive of later performance on tests of recall, recognition, and priming. *Society for Neuroscience Abstracts, 14,* in press.

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