SANDIA REPORT
SAND2012-8305
Unlimited Release
September 2012

Using Recordings of Brain Activity to Predict and Improve Human Performance

Laura E. Matzen, Michael J. Haass, Michael C. Trumbo, Austin Silva, Susan M. Stevens-Adams, Jennifer T. White, Anna Ho & David E. Peercy

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2012-8305
Unlimited Release
September 2012

Using Recordings of Brain Activity to Predict and Improve Human Performance

Laura E. Matzen, Michael J. Haass, Michael C. Trumbo, Austin Silva, Susan M. Stevens-Adams, Jennifer T. White, Anna Ho & David E. Peercy
Cognitive Systems Research and Applications
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS1327

Abstract

This project investigated methods for using recordings of event-related potentials (ERPs) to predict and improve human memory performance. In a series of experiments, the ERPs related to good and poor memory performance were characterized and a computational model was developed in which the ERPs recorded while a person studied words were used to predict whether that person would be a high performer or a low performer on a subsequent memory test. These experiments provide the first demonstration of ERPs used to predict subsequent memory performance. Additional experiments investigated the effects of different types of memory training (learning a memory strategy such as mental imagery or increasing memory capacity through working memory training) on memory performance and the associated brain activity.

ACKNOWLEDGMENTS

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was also made possible through the collaborations and support of Kara Federmeier at the University of Illinois at Urbana-Champaign, Michael Weisend at the Mind Research Network, and Michael Bunting, Isaiah Harbison, and Polly O'Rourke at the Center for the Advanced Study of Language. We also thank Laura Giffin, Sean Cusick, Bridget Milligan, and Charlotte Laguna for their assistance with data collection at the University of Illinois.

CONTENTS

1. Introduction 11
1.1. Background on Memory 11
1.1.1. Metamemory and Memory Strategies 12
1.1.2. Event-related Potentials 12
1.1.3. Modeling Event-related Potentials 13
1.1.4. Using Recordings of Brain Activity to Predict and Improve Human Performance 14

2. Experiment 1 – Characterizing Brain Activity Associated with Good and Poor Memory Performance 15
2.1. Method 15
2.1.1. Participants 15
2.1.2. Design 15
2.1.3. Materials 15
2.1.4. Procedure 16
2.1.5. Procedure for EEG Recording 17
2.2. Results 17
2.2.1 Behavioral results 17
2.2.2. Electrophysiological results 18
2.3. Discussion 23
2.3.1. Repeated words 23
2.3.2. Synonym and rhyme cues 24
2.3.3. The LPC and the benefits of retrieval practice 24

3. Experiment 2 – Effects of Retrieval Practice on Memory Performance 27
3.1. Methods 27
3.1.1. Participants 27
3.1.2. Materials 27
3.1.3. Procedure 28
3.2. Results and Discussion 28
3.2.1. Behavioral Results 28
3.2.2. ERP Results 29
3.2.3. Discussion 34

4. Computational Modeling of erps 35
4.1. Computational Modeling of Experiment 1 35
4.1.1. Modeling Results 36
4.2. Computational Modeling of Experiment 2 38
4.2.1 Modeling Results 38
4.3. Discussion 40

5: Experiment 3 - The Effects of Cognitive Training on Memory 42
5.1. Method 42
5.1.1. Participants 42
5.2. EEG Baseline Tasks 44
5.2.1. General Procedure 44

5.2.2. Procedure for EEG recording 44
5.2.3. Recognition Memory Task 44
5.2.4. Memory Error Task. 46
5.2.5. Associative Memory Task. 47
5.3. Behavioral Baseline Tasks 48
5.3.1. Deese-Roediger-McDermott (DRM) Task. 48
5.3.2 Listening Span Task. 49
5.3.3. Rotation Span Task. 50
5.4. Memory Training 51
5.4.1. Mental Imagery Training. 51
5.5. Working Memory Training 53
5.5.1. Adaptive N-back Task 53
5.5.2. Symmetry Span Task. 54
5.6. Behavioral Results 54
5.6.1. Mental Imagery Training 54
5.6.2. Working Memory Training 54
5.6.3. Baseline Tasks 55
5.7. Electrophysiological Results 61
5.7.1. Recognition Memory Task. 62
5.7.2. Associative Memory Task. ..... **Error! Bookmark not defined.**

6. References 70

Appendix A: development of Computational modeling methods for erps 74
A1.1. Data Overview 76
A1.2. Modeling Methods 77
A1.2.1. Euclidean Distance 77
A1.2.2. Statistical Distance 78
A1.2.3. Mahalanobis Distance 78
A1.2.4. Principal Component Analysis 78
A1.2.5. Independent Component Analysis 78
A1.2.6. Frequency Analysis 78
A1.2.7. Matched Filter 78
A1.2.8. Rational Model 79
A1.3. EEG Channel and Factor Selection 79
A1.3.1. Results 79
A1.4. References 81

Distribution 82

FIGURES

Figure 1. ERP waveforms for the first presentation and second presentations of repeated words. 19
Figure 2. ERP waveforms for the first presentation and second presentations of synonym pairs. 19
Figure 3. ERP waveforms for the first presentation and second 20
presentations of rhyming pairs. 20

Figure 4. ERP waveforms for the once-presented words that were subsequently remembered compared to those that were subsequently forgotten. 21
Figure 5. ERP waveforms illustrating the Dm (top row) and Dm-R (bottom row) effects for words repeated at short and long lags. 21
Figure 6. ERP waveforms illustrating the Dm (top row) and Dm-R (bottom row) effects for synonym pairs presented at short and long lags. 22
Figure 7. ERP waveforms illustrating the Dm (top row) and Dm-R (bottom row) effects for rhyming pairs presented at short and long lags. 22
Figure 8. Repetition Effect for Study-Study Short Lag Items. Black Line = First Presentation, Red Line = Second Presentation. 30
Figure 9. Repetition Effect for Study-Study Long Lag Items. Black Line = First Presentation, Red Line = Second Presentation. 30
Figure 10. Repetition Effect for Study-Test Short Lag Items. Black Line = First Presentation, Red Line = Second Presentation. 31
Figure 11. Repetition Effect for Study-Test Long Lag Items. Black Line = First Presentation, Red Line = Second Presentation. 31
Figure 12. Dm Effect for Once-Studied Words. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 32
Figure 13. Dm Effect for Study-Study Items, First Presentation. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 32
Figure 14. Dm Effect for Study-Study Second Presentation, Long Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 32
Figure 15. Dm Effect for Study-Study Second Presentation, Short Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 33
Figure 16. Dm Effect for Study-Test First Presentation. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 33
Figure 17. Dm Effect for Study-Test Second Presentation, Short Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 33
Figure 18. Dm Effect for Study-Test Second Presentation, Long Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items. 34
Figure 19. Proportion of "yes" responses for each test condition for the subject groupings created by the computational model. 37
Figure 20. Proportion of "yes" responses for each test condition for the subject groupings created by the computational model. 38
Fig. 21. Grand average ERPs to the unknown condition, the *second presentation of twice-studied words (black line)*, and the test condition, the *quizzed words (gray line)*. ERPs are shown at the midline central (MiCe) electrode. The participants whose brain activity was consistent with passive study in the unknown condition are shown on the left and the participants whose brain activity was consistent with a retrieval practice strategy are shown on the right. 39
Figure 22. Proportion of "yes" responses for each test condition for the subject groupings created by the computational model. 40
Figure 23. Distributions of age and educational background for the participants in the memory training study. 43
Figure 24. Average proportion of "yes" responses for each condition on the pre-training recognition memory test. 57

Figure 25. Difference between post-training memory performance (% correct) and pre-training memory performance for each condition on the recognition memory test, compared across training groups. 57
Figure 26. ERPs recorded at electrode Z7 for the first presentation of studied words and short- and long-lag repetitions..... 62
Figure 27. ERPs recorded at electrode Z7 for the first presentation of studied words and short- and long-lag quizzes. 63
Figure A1. SNR of each data ensemble normalized to the highest SNR ensemble..... 76
Figure A2. Subject average (dashed lines) and grand average (bold lines) ERP signals for FPSTW (red) & SPSTWSLR (black). 76
Figure A3. Subject average (dashed lines) and grand average (bold lines) ERP signals for S1R (red) & S1F (black)..... 77
Figure A4. Effect of number of subject used to form grand average ERP. The grand average from all 24 study subjects is shown in cyan and black. Grand averages formed by random selection of fewer than 24 subjects are shown in blue and red. (a. 24 random selections of 23 subjects to form grand average, b. 100 random selections of 15 subjects to form grand average, c. 100 random selections of 5 subjects to form grand average)..... 77
Figure A5. Illustration of time windows chosen for the rational model method. 79
Figure A6. Percentage of individual subject averages correctly classified in subject-out cross validation for each modeling and classification method..... 80
Figure A7. Percentage of individual subject averages correctly classified in subject-out cross validation for each modeling and classification method..... 80

TABLES

Table 1. Design of recall tests used in mental imagery training sessions..... 53
Table 2. Behavioral performance on recognition memory task..... 56
Table 3. Behavioral performance on memory error task 58
Table 4. Behavioral performance on associative memory task 59
Table 5. Behavioral performance on DRM task 60
Table 6. Performance on Rotation Span Task 61
Table 7. Performance on Rotation Span Task 61
Table A1. Categories of modeling methods 75

NOMENCLATURE

ACC	Anterior Cingulate Cortex
AUC	Area Under Receiver-operator Curve
Dm	Difference related to memory
DRM	Deese-Roediger-McDermott
EEG	Electroencephalogram
EOG	Electrooculogram
ERP	Event-Related Potential
ICA	Independent Components Analysis
ISI	Interstimulus Interval
KOhms	Kilo ohms
LDRD	Laboratory Directed Research and Development
LH	Left Hemisphere
LPC	Late Positive Component
MVSA	Multivariate Statistical Analysis
OSIQ	Object-Spatial Imagery Questionnaire
PCA	Principal Components Analysis
RH	Right Hemisphere
RMS	Root-mean-square
SNL	Sandia National Laboratories
SNR	Signal-to-Noise Ratio
VVIQ	Vividness of Visual Imagery Questionnaire

1. INTRODUCTION

In the LDRD "From Sensing to Enhancing Brain Processes," our aim was to improve human performance using brain activity recorded with electroencephalography (EEG). The focus of the project is on improving memory and decreasing memory errors for decision-relevant verbal/text-based information. To move toward the goal of improving memory, the overall LDRD included an experimental component and a computational modeling component. The experimental component of the project followed two interrelated tracks. The first track focused on characterizing brain activity associated with good and poor memory performance. The second track tested different methods for improving memory performance through cognitive training.

In the first and second years of the project, we conducted a series of experiments in which we characterized the event-related potentials (ERPs, electrophysiological responses observed during cognitive performance) associated with good and poor memory performance. We then used the ERPs to test different methods for developing computational models of the brain activity associated with good and poor memory performance. Using those models, we demonstrated that it is possible to use brain activity recorded while a person is studying information to predict whether or not that person will perform well on a later memory test. In the second and third years of the project, we conducted a memory training study in which participants' memory performance on a variety of tasks was assess before and after a three-week period of memory training. During the training period, one group learned a memory strategy (generating mental imagery as a memory aid), a second group engaged in working memory training in an effort to increase their working memory capacity, and a third group received no training and acted as a control group. This study provided novel information about the relationship between memory strategy and memory performance, validating the hypotheses generated by the modeling work.

1.1. Background on Memory

Memory underlies and supports all forms of high-level cognition and accurate memory is essential to good decision making. However, human memory is extremely fallible. Although there are many factors that can improve memory performance, such as selecting appropriate memory strategies, people are poor at predicting what they will or will not remember and tend to choose strategies that are suboptimal or counterproductive. In our research, we are investigating patterns of brain activity associated with good and poor memory performance. We are examining methods for improving human performance by identifying cases where learners are using suboptimal memory strategies. Through this effort, we hope to lay the foundation for closing the loop between recording brain activity and using those recordings to augment performance.

As a part of this effort, one of our goals is to create a model of brain activity that can be used in a predictive fashion. Using brain activity recorded from participants who tried to memorize words under a variety of study and test conditions, we selected two conditions where the brain's response to stimuli should be similar across all participants. We used those two conditions to develop a computation model and then tested the model on a third condition in which participants' brain activity should depend on their choice of study strategies. The model was used to predict which participants were using a more effective memory strategy. We then tested the predictions of the model by assessing the participants' behavioral memory performance.

1.1.1. Metamemory and Memory Strategies

The term *metamemory* refers to a person's judgments about the state of his or her own memory. Successful encoding and retrieval of information requires a number of metamemory decisions, such as deciding what information is worth remembering, what strategies should be used to encode the information, and whether or not information retrieved from memory is accurate. People typically develop metamemory skills over time, through experience with different kinds of learning situations. For example, after practice with sequences of study and test questions, people tend to get better at predicting which items they will remember later and which items need additional study (Benjamin, 2008). Through experience, people learn memory strategies such as spending more time studying items that seem difficult to remember or using different study strategies depending on when and how the information will need to be remembered. However, the strategies that learners develop are often affected by cognitive biases and may not be optimal. Numerous studies have shown that people often fail to use appropriate memory strategies (Benjamin, Bjork & Schwartz, 1998; Bjork, 1999; Koriat & Bjork, 2005; Koriat, Bjork, Sheffer & Bar, 2004; Schwartz, Benjamin & Bjork, 1997).

One memory strategy that can improve performance is self-testing, or retrieval practice (Allan & Rugg, 1997). A common example of retrieval practice is studying with flashcards. If a language student studies new vocabulary by quizzing herself with flashcards, she will be more likely to remember the new words than if she skimmed over the words and their definitions in a textbook. Retrieval practice is beneficial because it gives learners experience with retrieving the needed information from memory. It also provides learners with a more accurate sense of what they do and do not remember. The effectiveness of retrieval practice increases as the practice becomes more difficult (Bjork, 1999).

Although retrieval practice is a highly effective strategy, it is not a strategy that learners are likely to adopt on their own. Studying with this strategy can be frustrating because learners feel that they are performing poorly and progressing slowly. In reality, they are developing accurate assessments of how well they have learned the material. However, learners tend to prefer study strategies that make them feel successful at the time of study, even when those strategies are less effective in the long run (Benjamin, Bjork & Schwartz, 1998; Bjork, 1999).

1.1.2. Event-related Potentials

When a person's brain activity is recorded using electroencephalography (EEG), different patterns of activity emerge for passive study and for active retrieval of information from memory. In EEG research, a participant's brain activity is recorded using sensors placed on his or her scalp. The EEG data provide an ongoing record of the brain's electrical activity with very high temporal resolution. To separate out the brain activity related to a particular type of processing, the EEG data are time-locked to the presentation of events of interest and events of the same type are averaged together. The averaging process should average out any ongoing processing that was not related to the experimental stimuli, leaving only the activity that was elicited by the events of interest. These averaged waveforms are called event-related potentials (ERPs).

Researchers have mapped the relationships between different ERP waveforms and different types of processing in the brain. The ERP component of interest in the present study is the late positive component (LPC). The LPC is thought to be related to explicit processing, such as the process of deliberately searching memory for a particular piece of information (Neville, Kutas, Chesney & Schmidt, 1986; Paller & Kutas, 1992; Rugg & Doyle, 1994; Van Petten & Senkfor, 1996).

When a learner is presented with an item to study, the LPC elicited by that item will be small. However, when the learner is tested on an item and has to retrieve it from memory, the presentation of that item will elicit a large LPC. In the absence of an explicit memory test, a participant's self-induced retrieval practice should also produce a larger LPC. It is likely that use of retrieval practice as a memory strategy will be reflected in participants' brain activity during study.

1.1.3. Modeling Event-related Potentials

The EEG data selected for modeling was taken from a study in which participants were presented with a list of words and asked to remember them for a later memory test. Some of the words were studied once, some words were studied twice, and some were studied once and then quizzed once during the study session. All of the words appeared again on a subsequent memory test, intermixed with an equal number of new words. We hypothesized that participants would have the worst memory for the words that were studied only once and the best memory for the words that were quizzed during the study sessions. The quizzes provide an opportunity for retrieval practice that should benefit subsequent memory performance.
For the words that were studied twice but not quizzed, we hypothesized that some participants would engage in retrieval practice on their own. Even though the words were not explicitly tested, participants might recognize them as previously studied words and retrieve the first presentation of the word from memory. This self-testing should benefit subsequent memory performance much like explicit testing. Since, as discussed above, most people are unlikely to adopt a strategy such as retrieval practice on their own, we expected that the average performance across all participants would be lower for the twice-studied items than for the quizzed items. However, we expected that a subset of the participants would use more effective memory strategies and would perform better on this condition than their peers.

The design of the EEG experiment allowed us to model each participant's brain activity in two "known" conditions: the first presentation of each studied word, which should not elicit an LPC, and the words that were quizzed during the study block, which should elicit a large LPC. We applied the model to ERPs from an unknown condition, the second presentation of repeated study words. The words in that condition should elicit an LPC only for the participants who engaged in retrieval practice. We used the model to classify the ERPs from the unknown condition as being more like passively studied words or more like explicitly tested words. We then tested the predictive power of the model by comparing the subsequent memory performance for participants in those two groups.

1.1.4. Using Recordings of Brain Activity to Predict and Improve Human Performance

In this report, we describe the results of three ERP experiments intended to lay the framework for designing interventions that can apply recordings of brain activity to improving human performance. Experiments 1 and 2 focused on characterizing patterns of brain activity associated with good and poor subsequent memory performance. The data from those experiments was then used to develop and test a computational model that is able to predict a person's subsequent memory performance. Experiment 3 built upon the earlier work by testing different methods for improving memory performance through cognitive training. The results of Experiment 3 suggest that there is a trade-off between different types of memory performance and that training should be carefully tailored to the needs of an individual learner.

2. EXPERIMENT 1 – CHARACTERIZING BRAIN ACTIVITY ASSOCIATED WITH GOOD AND POOR MEMORY PERFORMANCE

The goal of this experiment was to characterize the brain activity associated with good and poor memory performance. In this case, good memory performance was simply remembering an item later, while poor memory performance was forgetting. Unlike traditional Dm Effect experiments, in which participants study a word once and are tested on it once, we used multiple study and test combinations in order to make the experiment more like a real-world memory task. In real-world tasks, if people are trying to learn or remember some information, they are likely to study that information multiple times. They may also be tested on the information multiple times and in multiple settings. In Experiment 1, our aim was to use a variety of study and test conditions to gain more insight into the origin and nature of the Dm Effect.

2.1. Method

2.1.1. Participants

Thirty-seven University of Illinois students participated in the experiment for credit in an introductory psychology course or for payment. All were right-handed, had no early exposure to languages other than English, and had no history of neurological disease or defect. Seven participants were excluded from the final data analysis due to exceptionally poor performance on the memory test. The remaining thirty participants (17 female) were included in the ERP data analysis. The mean age of these participants was 20 (range 18-28).

2.1.2. Design

The critical variables in Experiment 2 were the types of items that were associated with each parent word (the same word presented again, a synonym, a rhyming word, or no associated item) and the lag between the parent word and its associated item (one or nine intervening items). Both of these variables were manipulated within-subjects. Participants were tested on all of the parent words, synonyms and rhyme words along with an equal number of new, unrelated words in a recognition memory test. The dependent variable for the behavioral test was yes/no recognition for the parent words in each condition at each lag.

2.1.3. Materials

Experiment 2 used a list of 320 parent words, each of which had a synonym and a rhyme. For example, one parent word is “cloth,” which can be paired with its synonym, “fabric,” or with a rhyming word such as “moth”. The items were divided into seven experimental lists. Within each list, 80 of the parent words were presented only once, 80 were presented twice, 80 were presented with their synonyms and 80 were presented with their rhymes. In the twice presented, synonym, and rhyme conditions, half of the items were in the short lag condition, with one intervening item between the parent word and its reminding cue. The other half of the items were presented in the long lag condition, with nine intervening items between the parent word

and its reminding cue. This design gives rise to seven conditions. The items in each condition were counterbalanced across lists.

A set of 480 new, unrelated words was created for use in the recognition test. Of the new words, 320 were matched to the parent items on length, frequency and concreteness. None of these words were synonyms or rhymes of any parent word. An additional 160 new words were matched to the synonym and rhyme cues on length, frequency, and concreteness. Once again, none of these words were synonyms of any of the parent words. However, words that rhyme with one or more parent words were included in this set because the words could not otherwise be matched on length and frequency. In addition, including words that might rhyme with others in the list made this set of new items more comparable to the old test items, half of which rhymed with previously studied words.

Within each experimental list the words were assigned to a pseudorandom order with the appropriate lag structures and with no more than five items from the same condition in sequence. As in Experiment 1, the lists were divided into four blocks to make the memory task easier for the participants. Each block contained 140 study words, including 80 parent words, 20 synonyms, 20 rhymes, and 20 re-presentations of parent words. Each study block was followed by a recognition test containing 240 test words. The first 160 test words included the 80 parent words from the preceding study block and 80 matched new words. The next 80 test words included the 20 synonyms, the 20 rhymes, and 40 matched new words. The memory test for the reminding cues followed the memory test for the parent words in order to avoid any reminding that might occur at the time of test rather than at the time of study. However, from the point of view of the participant, there was no difference between the two parts of the test block.

2.1.4. Procedure

The participants were instructed that they would be tested on their memory for a list of study words. They were not informed of the presence of synonyms, rhymes, or repeated words within the study list. Throughout the experiment, there was a white fixation cross in the center of the computer screen. The participants were asked to keep their eyes on the fixation cross at all times during the experiment. All of the study and test words were presented immediately above the fixation cross in white 38-point Helvetica font on a black background. Within the study blocks, each word was preceded by a red dot that was presented above the fixation cross for one second. Participants were instructed that they could blink or move their eyes while the red dot was on the screen, but that when it disappeared they should refrain from blinking and prepare to see the next study word. The study word was presented 500 ms after the red dot disappeared and remained on the screen for one second. During the test blocks, the test words appeared on the screen for one second, followed by a question mark. Participants were instructed to indicate whether or not they remembered studying the preceding word. They responded by pressing one of two response buttons, labeled “yes” and “no”. In either case, the question mark remained on the screen until the participant makes a response. Once the participant had responded, a blank screen was presented for one second, followed by the presentation of another test word. The participants took short breaks before starting each new study block in order to reduce interference from the preceding blocks.

2.1.5. Procedure for EEG Recording

The electroencephalogram (EEG) was recorded from 26 silver/silver-chloride electrodes embedded in a geodesic arrangement in an elastic cap (EASY-cap). The locations of the 26 electrodes were midline prefrontal (MiPf), left and right medial prefrontal (LMPf and RMPf), left and right lateral prefrontal (LLPf and RLPf), left and right medial frontal (LMFr and RMFr), left and right mediolateral frontal (LDFr and RDFr), left and right lateral frontal (LLFr and RLFr), midline central (MiCe), left and right medial central (LMCe and RMCe), left and right mediolateral central (LDCe and RDCe), midline parietal (MiPa), left and right mediolateral parietal (LDPa and RDPa), left and right lateral temporal (LLTe and RLTe), midline occipital (MiOc), left and right medial occipital (LMOc and RMOc), and left and right lateral occipital (LLOc and RLOc). Five additional free electrodes were placed on the left and right mastoids, on the outer canthus of each eye, and below the left eye. The three free electrodes near the eyes were used to record blinks and horizontal eye movements (vertical and horizontal EOG). The scalp electrodes were referenced on-line to the left mastoid. Following the experiment, the scalp electrodes were re-referenced off-line to an average of the left and right mastoids. All of the electrodes were tested before recording begins to ensure that their impedance was below 3 KOhms. During the experiment, the EEG from all electrodes was amplified through a bandpass filter of 0.02-100 Hz and recorded at a sampling rate of 250 Hz.

ERPs were computed at each electrode for each experimental condition by averaging the EEG data from 100 ms before the onset of a word until 920 ms after word onset. Trials containing blinks were corrected using the blink correction procedure described by Dale (1994) and trials containing artifacts such as excessive eye movement, signal drift or muscle activity were excluded from the averages. The mean amplitude of the ERPs within time windows of interest was calculated using data digitally filtered off-line using a bandpass filter of 0.2 to 20 Hz.

2.2. Results

2.2.1 Behavioral results

On average, the participants correctly rejected the new items 82% of the time and correctly endorsed the old items 65% of the time. In order to evaluate the effects of the reminding cues, memory performance for each of the reminding conditions was compared to the baseline of memory performance for words that were presented only once during study, with no associated reminding cue. On average, the participants correctly recognized 61% of the once-presented words. As predicted, their memory performance was significantly better for the words that were presented twice, with hit rates of 70% for words that were repeated at short lags ($t(29) = 5.26$, $p < 0.001$) and hit rates of 73% for words that were repeated at long lags ($t(29) = 6.93$, $p < 0.001$). The spacing effect was also evident in the twice-presented condition, with memory performance being reliably higher for long lag repetitions than for short lag repetitions ($t(29) = 1.90$, $p = 0.06$).

In the synonym condition, memory performance was not significantly higher than it was in the once-presented condition for either the short lags (62%, $t(29) = 0.55$) or the long lags (63%, $t(29) = 1.65$). In addition, there was no spacing effect for the synonym condition. Performance in the

long lag condition was not significantly higher than it was for the short lag condition ($t(29) = 0.68$). The same was true for the rhyme condition. The average hit rate was 61% for both the short and long lag conditions, which was identical to the average hit rate for the once-presented words.

The behavioral results indicate that the synonyms and rhymes were not effective reminding cues. Unlike the repetitions, the presentation of a synonym or rhyme related to a parent word did not provide any benefit to memory for the parent word.

2.2.2. Electrophysiological results

Within the grand average waveforms, the N400 was measured by computing the mean amplitude of the ERPs in a time window from 300-500 ms post stimulus onset. The LPC was measured by computing the mean amplitude of the ERPs in a time window from 500-900 ms post stimulus onset. The Dm effect (for the parent words) and Dm-R effect (for the repetitions of parent words) were measured by computing the mean amplitude of the ERPs for remembered words versus forgotten words in both the 300-500 ms and in the 500-900 ms time windows. Repeated measures ANOVAs were used to test the results, with degrees of freedom adjusted using the Greenhouse-Geisser correction.

2.2.2.1. N400 repetition effect

Figures 1, 2, and 3 show the ERPs elicited at study for the critical words and reminding cues in the twice-presented, synonym, and rhyme conditions, respectively. During the study phase of the experiment, there was a significant reduction in the amplitude of the N400 component for the second presentation of the twice-presented words relative to the first presentation at short lags ($F(1,29) = 12.40$, $p < 0.01$) but not at long lags ($F(1,29) = 0.19$). These data indicate that there was an N400 repetition effect for repeated words at short lags, but that the effect was attenuated at long lags.

In the synonym and rhyme condition, there was no significant reduction in the amplitude of the N400 at either lag (all F's < 1.05). These data indicate that there was no N400 repetition effect for the synonym and rhyme cues.

2.2.2.2. Late positive component

For the twice-presented condition, there was a significant difference in the amplitude of waveform in the LPC time window for both the short lag repetitions ($F(1,29) = 16.88$, $p < 0.001$) and the long lag repetitions ($F(1,29) = 6.81$, $p = 0.02$) relative to the first presentation of the words. A comparison of the LPCs at the short and long lag repetitions revealed that the amplitude of the LPC was significantly larger for short lag repetitions ($F(1,29) = 4.07$, $p = 0.05$).

For the synonym and rhyme conditions, there were no significant differences in the amplitudes of the LPCs for either the short or long lag conditions relative to the presentation of the critical words (all Fs < 2.6, all p's > 0.11).

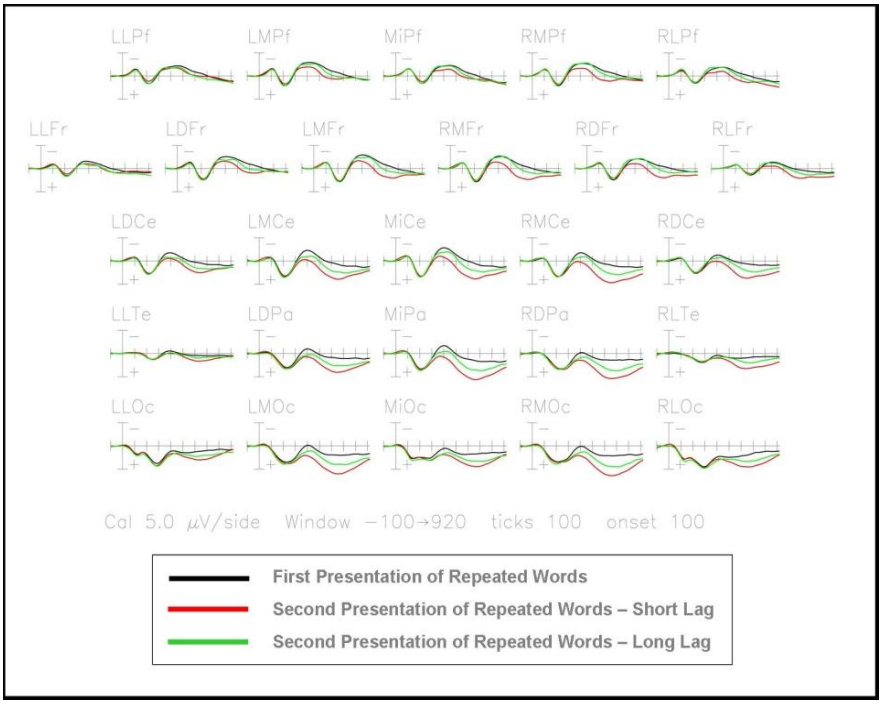


Figure 1. ERP waveforms for the first presentation and second presentations of repeated words.

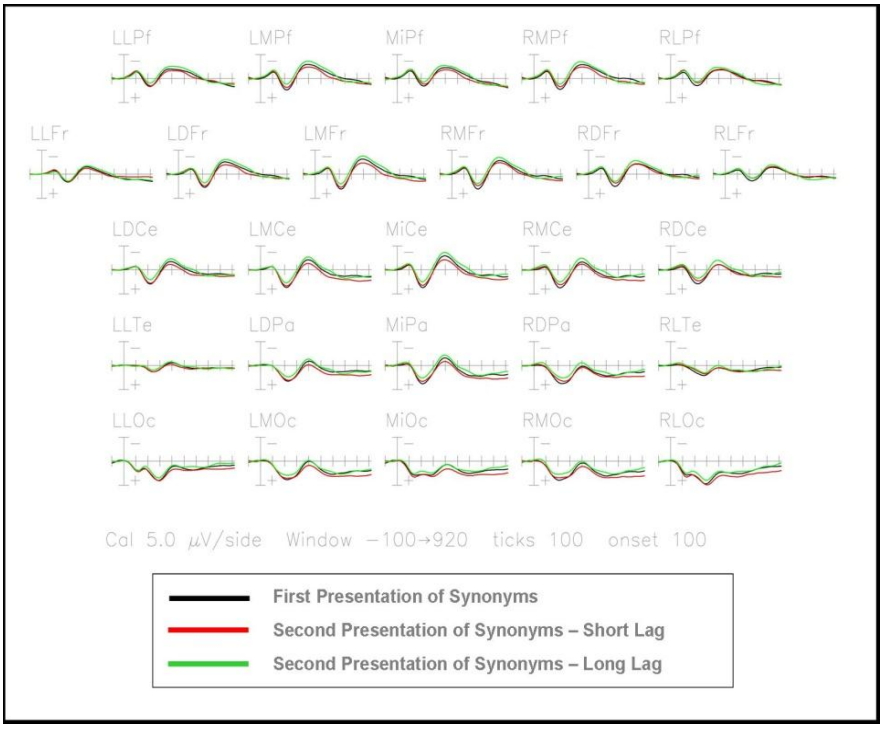


Figure 2. ERP waveforms for the first presentation and second presentations of synonym pairs.

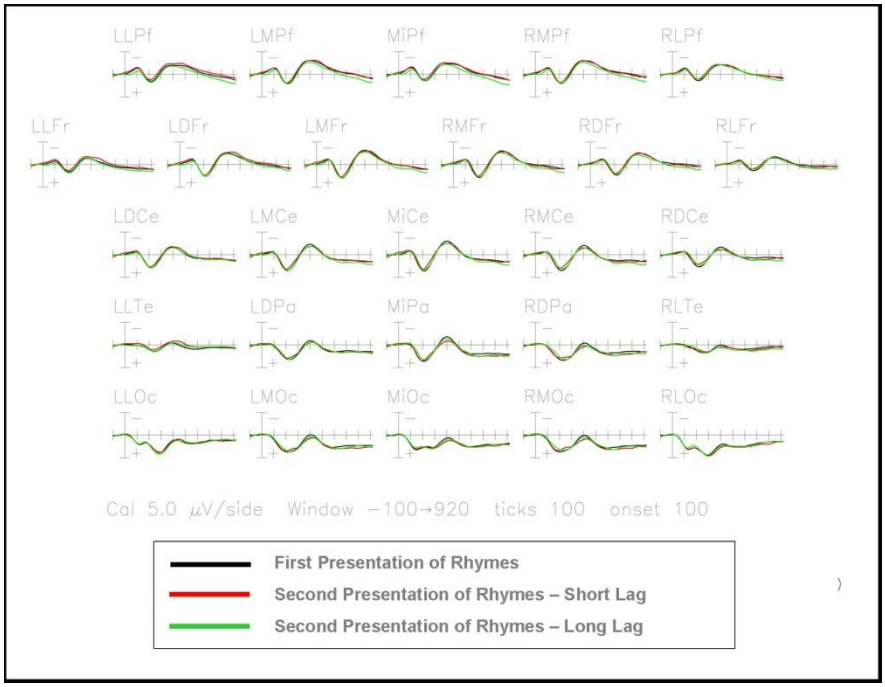


Figure 3. ERP waveforms for the first presentation and second presentations of rhyming pairs.

2.2.2.3. Dm effect

The traditional Dm effect was calculated by backsorting the ERPs elicited by the critical words based on subsequent memory for those words. For all conditions the Dm effect was largest over posterior channels, so analyses of the Dm effect were carried out using data from the nine electrodes showing the largest Dm effects across conditions: LMCe, MiCe, RMCe, LDPa, MiPa, RDPa, LMOc, MiOc, and RMOc.

For the once-presented words, the ERPs for the subsequently remembered words were significantly more positive than the ERPs for the subsequently forgotten words only in the 200-350 ms time range ($F(1,29) = 4.33$, $p = 0.05$). Figure 4 shows the backsorted ERPs for the once-presented condition.

For the twice-presented, synonym, and rhyme conditions, the Dm effect did not reach significance in any time window (all F's < 2.67). The top rows of Figures 5, 6 and 7 show the backsorted ERPs at each lag for the twice-presented, synonym and rhyme conditions, respectively.

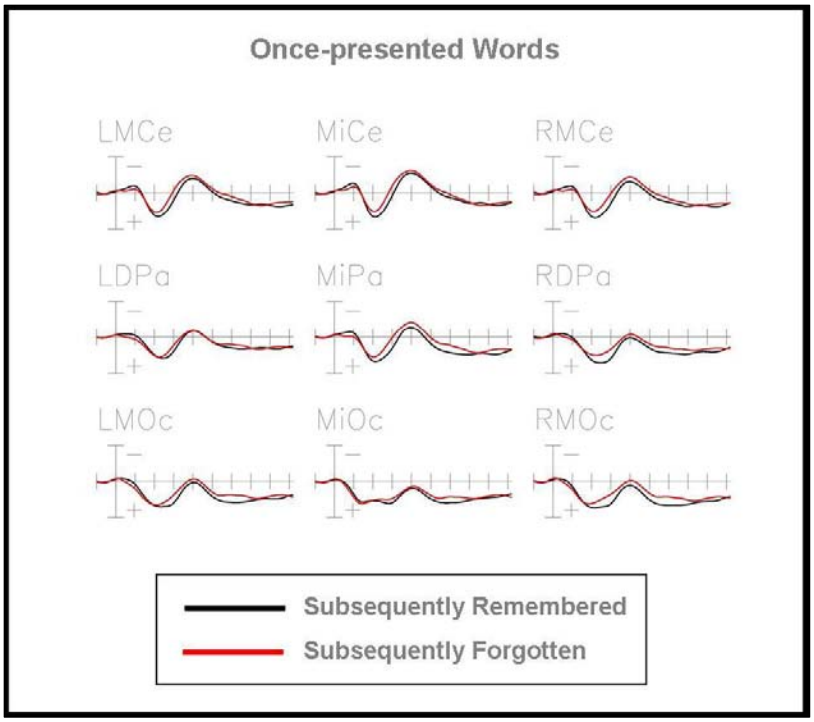


Figure 4. ERP waveforms for the once-presented words that were subsequently remembered compared to those that were subsequently forgotten.

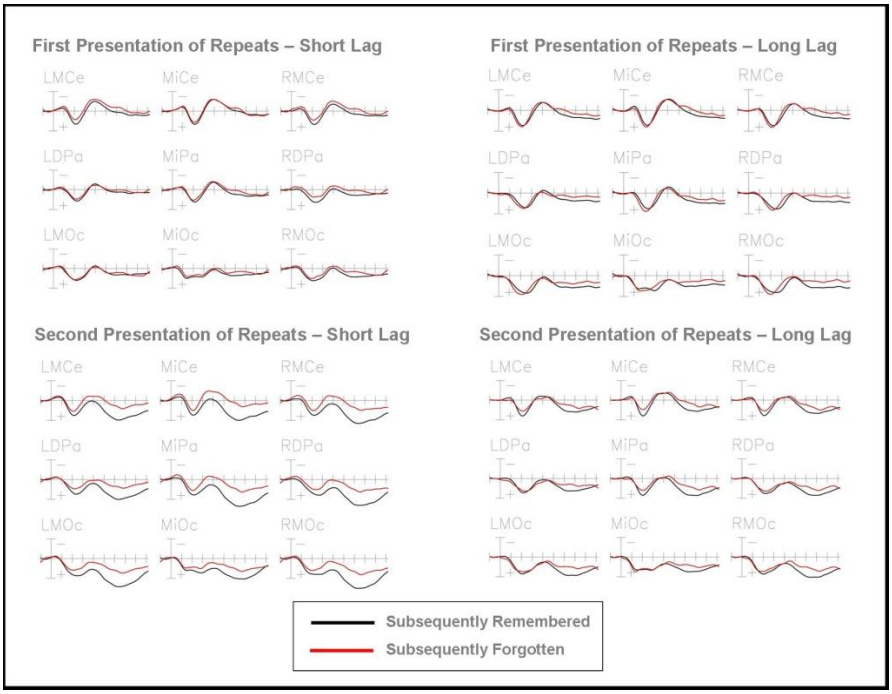


Figure 5. ERP waveforms illustrating the Dm (top row) and Dm-R (bottom row) effects for words repeated at short and long lags.

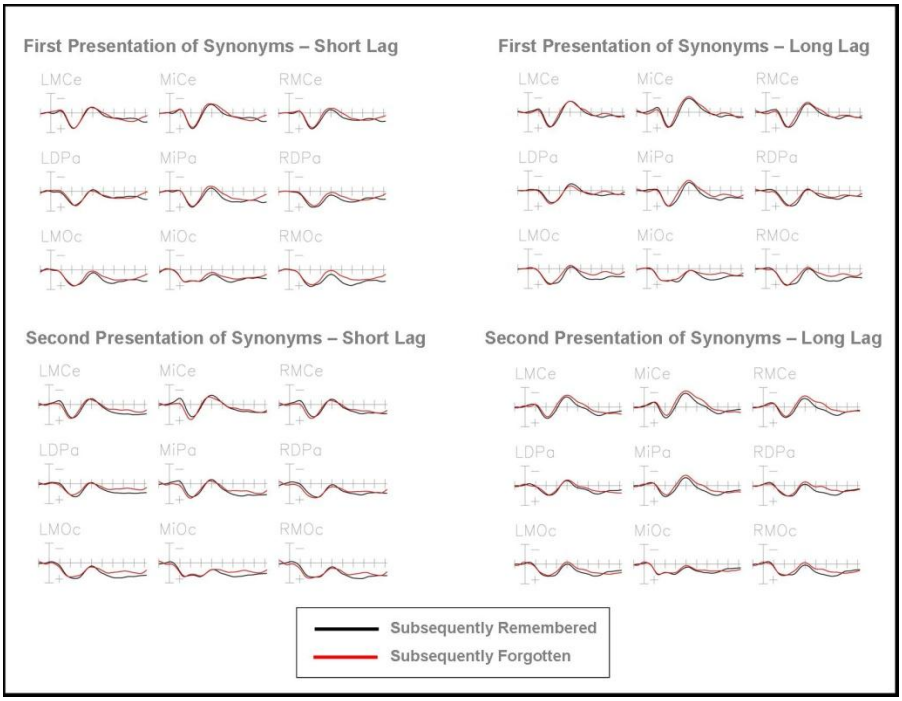


Figure 6. ERP waveforms illustrating the Dm (top row) and Dm-R (bottom row) effects for synonym pairs presented at short and long lags.

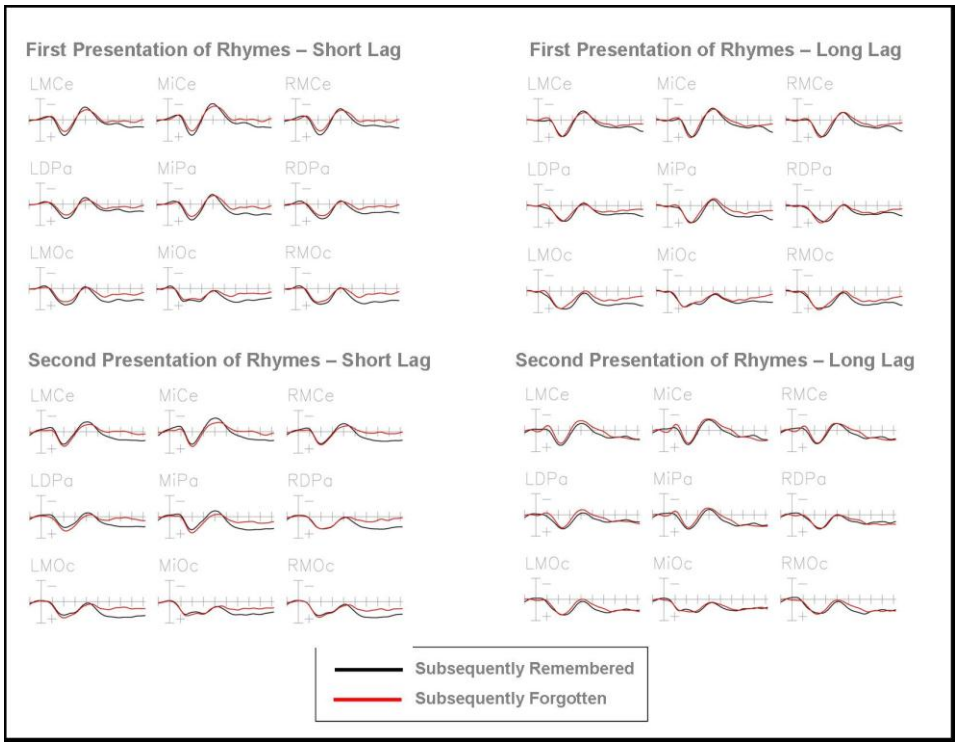


Figure 7. ERP waveforms illustrating the Dm (top row) and Dm-R (bottom row) effects for rhyming pairs presented at short and long lags.

2.2.2.4. Dm-R effect

In order to study the impacts of the repetitions on subsequent memory, Dm effects were also calculated for the second presentation of repeated words. This effect was termed the Dm-Repetition, or Dm-R effect. The Dm-R effect was calculated by backsorting the ERPs elicited by the second presentation of repeated words based on subsequent memory for those words. As with the standard Dm effect, the analyses for the Dm-R effect were carried out using data from electrodes LMCe, MiCe, RMCe, LDPa, MiPa, RDPa, LMOc, MiOc, and RMOc. Analyses of the mean amplitude of the waveforms were carried out in the N400 time window from 300-500 ms and in the LPC time window from 500-800 ms. The bottom rows of Figures 5, 6 and 7 show the Dm-R effect for the repeated words, synonyms and rhymes at each lag.

In the N400 time window, there was a significant Dm-R effect elicited by the second presentation of twice-presented words at short lags ($F(1,29) = 7.49$, $p = 0.01$), but not for long lags ($F < 1.00$). In the LPC time window, the Dm-R effect was significant at both the short ($F(1,29) = 14.51$, $p < 0.001$) and long ($F(1, 29) = 4.44$, $p < 0.04$) lags.

2.3. Discussion

The behavioral results from Experiment 1 revealed that, as predicted, seeing the same word twice during study led to improved memory at test. There was also a spacing effect for the repeated words, with better performance for words repeated at long lags than for words repeated at short lags. However, in this experiment, seeing a synonym or rhyme of a critical word during study did not lead to improved memory at test or a spacing effect. Memory performance in those conditions was the same as the performance for the once-presented words that had no associated reminding cue. These results indicate that synonyms and rhymes were not effective reminding cues under the study conditions used in this experiment.

2.3.1. Repeated words

The ERP data for the words that were repeated during the study session showed an N400 repetition effect for words repeated at short lags and a large LPC for repetitions at both short and long lags, with a larger amplitude LPC at short lags. The smaller N400 for repeated words at short lags indicates facilitation or priming for the second presentation of the word and is consistent with previous studies (Anderson & Holcomb, 2005; Bentin & Peled, 1990; Besson & Kutas, 1993; Besson et al., 1992; Nagy & Rugg, 1987; Olichney et al., 2000; Rugg, 1985; 1987; 1990; Smith & Halgren, 1987, Van Petten, et al., 1991). At long lags, the N400 repetition effect no longer reached significance. This is consistent with previous studies that have found that N400 facilitation for repeated words fades fairly rapidly in word lists (Bentin & Moscovitch, 1990; Rugg & Doyle, 1994). In the present experiment, the onset of the lag 2 repetitions occurred four seconds after the offset of the first presentation of the word, but the lag 10 repetitions fell well outside of this time range.

The LPC results for the repeated words are consistent with the idea that the LPC can be used as an index of reminding. As discussed previously, LPCs are not typically seen in studies using implicit repetition during study (Laszlo & Federmeier, 2007). Rather, they are elicited by words

that lead to explicit processing (Allan & Rugg, 1997; Olichney et al, 2000; Paller & Gross, 1998; Paller & Kutas, 1992; Rugg & Doyle, 1994; Van Petten & Senkfor, 1996), which are commonly test items. The presence of an LPC in response to the repetitions the present experiment indicates that the participants are consciously remembering having seen that word earlier. At short lags, it is highly likely that participants will realize that they have just seen the same word twice, leading to a large LPC. At long lags, there should still be a large LPC for words that trigger reminding for the previous occurrence of the same word. However, at long lags it is also more likely that the participant could have forgotten the first instance of the word and will fail to retrieve it. In these cases, an LPC will not be elicited. Thus, on average, the amplitude of the LPC for long lag repetitions is smaller than it is for short lag repetitions.

2.3.2. Synonym and rhyme cues

This experiment did not find N400 repetition effects or larger LPCs in response to synonym or rhyme cues, likely because the participants were not alerted to the presence of synonyms and rhymes in the study list. It was somewhat surprising that there was no N400 repetition effect for the synonyms presented at short lags. The N400 repetition effect has been shown to be related to semantic priming or semantic facilitation in many previous studies (Anderson & Holcomb, 2005; Kutas & Federmeier, 2000) and we expected that this facilitation would occur for the synonyms in this experiment. It is possible that the presence of repeated words within the study list and the fact that the participants were not informed of any possible relationships between the studied words weakened the effects of the semantic overlap between the synonym pairs. There might have been more N400 facilitation for the synonyms if the participants' attention had been called to the patterns within the study lists.

Similarly, there were no LPCs elicited by synonyms or rhymes at either short or long lags, indicating that these conditions did not lead to explicit processing such as recalling previously studied words. Once again, if the participants were given a task that called their attention to the semantic or phonological relationships between words, that may have made the synonyms or rhymes more likely to lead to the recollection of previously studied words. However, this experiment shows that when participants are passively reading words and are not informed about possible relationships between words in the list, synonyms and rhymes are not effective reminding cues.

2.3.3. The LPC and the benefits of retrieval practice

One of the goals of this experiment was to use the amplitude of the LPC as an index of explicit reminding or retrieval practice. The combination of the LPC data and the behavioral data indicates that the conditions that led to the best memory performance were the same conditions that were most likely to elicit an LPC. The twice-presented condition was the only one that led to improved performance at test, and was also the only condition that elicited LPCs. This pattern indicates that the presence of an LPC at study is associated with improved memory performance at test. This relationship can also be tested directly by examining the Dm-R effects for reminding cues when they are sorted based on subsequent memory for their associated critical words.

The Dm-R effects in this experiment revealed that the amplitude of the LPC elicited by effective reminding cues at test is predictive of subsequent memory for their associated critical words. In this experiment, only the repeated words acted as effective reminding cues. While the traditional Dm effect for the first presentations of the repeated words looks very similar to the Dm effects seen for once-presented words, the Dm-R elicited by the second presentation of repeated items looks quite different. For repeated words presented at short lags, the Dm-R is very large and spans both the N400 and LPC time windows. For repeated words presented at long lags, the Dm-R effect is also large, but it appears primarily in the LPC time window. In both cases, the second presentation of words that are later remembered have much larger LPCs than words that are later forgotten. This pattern provides direct evidence for a relationship between the amplitude of the LPC in response to a reminding cue and subsequent memory performance at test

Taken together, the results of this experiment support the hypothesis that reminding during study plays an important role in producing the spacing effect. The prior literature on the spacing effect suggests that when study words are repeated in close proximity to one another, the repetition provides little benefit to memory performance. As the lag between the two presentations increases, the second presentation is more likely to improve memory performance, either because it reminds participants of the first occurrence of the word, reinforcing memory for that instance, or because participants are more likely to encode the second occurrence of the word as a separate memory trace, doubling their chances of recalling that word. If reminding drives the spacing effect, the ERP components associated with explicit processing would be more predictive of later memory performance than the components associated with more automatic processes such as facilitation of semantic or phonological access. As predicted, this was the pattern that was found in Experiment 1. For the twice-presented words, the only condition that produced a spacing effect, the amplitude of the LPC elicited by the reminding cues was predictive of subsequent memory. This indicates that the words that led to explicit processing during study, which in this case we believe to be retrieval of previously studied words, are the ones that are most likely to be remembered at test.

If the spacing effect is driven by explicit reminding, as suggested here, the process of reminding provides the biggest benefit to memory when there is a long lag between two presentations of the same word. This indicates that reminding provides the biggest benefit when it is more difficult. We hypothesized that the process of retrieving a previously encoded word during study reinforces the encoding of that word or provides practice with retrieval that is beneficial at test. As it becomes more difficult to retrieve the word, as when there is a long lag between repetitions, successful retrieval becomes increasingly beneficial. Conversely, when there is a short lag between repetitions and semantic access is facilitated by implicit priming, retrieving the word is easier and there is less of a benefit to subsequent memory performance. The data from Experiment 1 are consistent with the view that conditions with explicit reminding lead to improved memory performance, but that there is less of an improvement in conditions that also show evidence for implicit priming.

In summary, Experiment 1 revealed that the LPC is associated with explicit reminding, and that the presence of an LPC at study is predictive of improved memory performance at test. This supports the hypothesis that the spacing effect is driven by reminding occurring during the study

phase. Experiment 1 did not provide evidence suggesting that either synonyms or rhymes were effective as reminding cues. Rather, the experiment showed that when participants were unaware of the possible relationships between words, both semantic and surface form overlap were needed to trigger reminding. There are several possible reasons for this outcome. It is possible that when there are strong reminding cues available, people are less likely to use words with weaker relationships to one another as reminding cues. Additionally, the participants most likely treated the synonym and rhyme cues as additional words that they needed to remember, rather than using them as cues to retrieve previously studied items. In future experiments, it would be useful to investigate this issue further by alerting participants to the relationships between words or to the status of synonyms and rhymes as reminders rather than additional words to be remembered. It may also be beneficial to give the participants specific task during the study phase that requires them to relate the words to one another. This would provide additional information about the circumstances under which semantic and surface form similarities might become useful reminding cues.

3. EXPERIMENT 2 – EFFECTS OF RETRIEVAL PRACTICE ON MEMORY PERFORMANCE

In Experiment 2, we sought to extend the findings of Experiment 1 by explicitly quizzing participants on studied words during the study session. If our conclusions from Experiment 1 were correct, quizzing should provide an even larger benefit to memory performance than repeating studied words. By explicitly asking participants to search their memory for previously studied items, the effect of retrieval practice should be strengthened. In addition, if repeated study words elicit a large LPC because they remind participants of previously studied words, there should also be a large LPC elicited by quizzed words as participants attempt to recollect previously studied items.

3.1. Methods

3.1.1. Participants

This experiment was conducted in Dr. Kara Federmeier's lab at the University of Illinois at Urbana-Champaign. A total of 30 people participated in the experiment. All were right-handed, had no early exposure to languages other than English, and had no history of neurological disease or defect. Six were excluded from the data analysis due to excessive artifacts (eye blinks) in the EEG data. Of the remaining 24 participants who were included in the analysis, 12 were male and 12 were female. The average age of the participants was 21.

3.1.2. Materials

The materials used in the experiment consisted of 320 common nouns that served as study items, and 320 nouns that were matched in terms of length and frequency and served as new items at test. The average frequency was 57.6 for the study items and 50.9 for the new items; the average word length was 4.6 letters for both sets of words (frequency data was taken from the Kucera and Francis, 1967; norms included in Balota et al., 2002; a frequency value of zero was assumed for items not appearing in the database).

The study words were divided into eight counterbalanced lists. The experimental lists were subdivided into four study blocks and four test blocks. Each study block contained 60 of the experimental items. Of those items, 20 were studied once, 20 were studied twice, and 20 were studied and then tested within the block.[1] For the items that were studied twice or studied and then tested, half of the items were repeated at a short lag, defined as one intervening item, and half were repeated at a long lag, defined as nine intervening items.

The study words were divided into eight counterbalanced lists. The experimental lists were subdivided into four study blocks and four test blocks. Each study block contained 80 of the experimental items. Of those items, 20 were studied once, 20 were studied twice, 20 were

[1] Each study block also contained 20 nouns that were paired with a synonym. The synonym was presented either at a short or long lag, and half of the synonyms were tested. These items were treated as filler items and are not included in the analysis discussed below.

studied and then tested within the block, and 20 were paired with a synonym. For the items that were studied twice or studied and then tested, half of the items were repeated at a short lag, defined as one intervening item, and half were repeated at a long lag, defined as nine intervening items. For the items that were paired with synonyms, half of the synonyms were presented at a short lag and half were presented at a long lag. In addition, half of the synonym items were tested at each lag.

Each study block was followed by a test block in which all of the nouns from the block were tested or re-tested, intermixed with an equal number of new, unstudied items.

3.1.3. Procedure

The participants were instructed that they would be tested on their memory for a list of study words. They were not given any information about different types of memory strategies and were not asked to use a particular memory strategy. As discussed above, the experiment was broken into four study-test blocks in order to make the task easier for the participants. Each study block contained a total of 140 study words and each test block contained a total of 160 test words.

Throughout the experiment, there was a white fixation cross in the center of the computer screen. The participants were asked to keep their eyes on the fixation cross at all times during the experiment. All of the study words were presented immediately above the fixation cross in white 38-point Helvetica font on a black background. Within the study blocks, each word was preceded by a pound symbol (#) that was presented above the fixation cross for one second. Participants were instructed that they could blink or move their eyes while the pound symbol was on the screen, but that when it disappeared they should refrain from blinking and prepare to see the next study word. For the tested words, the pound symbol was red, indicating that the next word would be tested. For the words that were only studied, the pound symbol was white. The study and test words were presented 500 ms after the pound symbol disappeared and remained on the screen for one second. The tested words were followed by a red question mark that remained on the screen until the participants pressed a response button to indicate whether or not that word had appeared earlier in the study block. The same procedure was used in the test blocks that followed each study block. In the test blocks, all of the words from the study block were tested or retested, intermixed with an equal number of new words. The participants took short breaks before starting each new study block in order to reduce interference from the preceding blocks.

The procedures used for the ERP recording in Experiment 2 were identical to the procedures used in Experiment 1.

3.2. Results and Discussion

3.2.1. Behavioral Results

Memory accuracy was assessed using the percentage of correct answers on the memory tests. For the quizzes that occurred during the study session, the participants were on average 88% correct for quizzes that occurred at a short lag after the corresponding study item and 55% correct for

quizzes that occurred after a long lag. The participants correctly rejected 92% of the new, unstudied words that were quizzed during the study phase. The quizzed words that were synonyms of studied words were also rejected 92% of the time at both short and long lags.

During the memory tests that followed the study blocks, participants were on average 39% correct for words that were studied only once, 49% correct for words that were studied twice with a short lag between repetitions, 50% correct for items that were studied twice with a long lag between the repetitions, 62% correct for the items that were studied once and then quizzed after a short lag, and 66% correct for items that were studied once and quizzed after a long lag. The participants correctly rejected new, unstudied items presented at test 86% of the time. For words that had a synonym presented during study, the participants' performance was similar to their performance for words that were only studied once, with average scores of 40% correct for both the short lag and the long lag condition. When the synonym was quizzed after a short lag, performance for the parent item was slightly higher at 50% correct. When a synonym was quizzed after a long lag, memory for the parent item dropped down to 43% correct.

These results were consistent with the prediction that retrieval practice during study would benefit subsequent memory performance. The difference in performance between the twice-studied words and the quizzed words also supports the hypothesis that most participants would not use retrieval practice when presented with repeated study words.

3.2.2. ERP Results

We focused on two time windows of interest within the grand average waveforms: the N400 and LPC time windows. The N400 was measured by computing the mean amplitude of the ERPs in a time window from 300-500 ms post stimulus onset. The LPC was measured by computing the mean amplitude of the ERPs in a time window from 500-900 ms post stimulus onset. Similarly, the Dm Effect was measured by computing the mean amplitude of the ERPs for remembered words versus forgotten words in the 300-500 ms and 500-900 ms time windows. Repeated measures ANOVAs were used to test the results, with degrees of freedom adjusted using the Greenhouse-Geisser correction. All effects are significant at or above the $p = 0.05$ level unless otherwise specified.

3.2.2.1. N400 Repetition Effect

Figures 8, 9, 10 and 11 show the ERPs elicited at study for the critical words and reminding cues in the study-study short lag, study-study long lag, study-test short lag and study-test long lag, respectively. During the study phase of the experiment, there was a significant reduction in the amplitude of the N400 component for the second presentation of the study-study short lag words and for the study-test words at both lags. These data indicate that there was an N400 repetition effect for repeated words at short lags, but that the effect was attenuated at long lags. However, when participants know that they will be tested on the word, they may think of recently studied words, leading to N400 facilitation.

3.2.2.2. Late Positive Component

As with the N400, there was a significant difference in the amplitude of waveform in the LPC time window for the second presentation of study-study short lag items, study-test short lag items, and study-test long lag items relative to the first presentation. The LPC is particularly large for the tested items, indicating that participants were actively searching their memory for the test word, as predicted. For the study-study short lag items, the LPC likely indicates that the participants recognize that they have just studied the same word. However, for the study-study long lag words, the absence of an LPC indicates that participants are not actively searching their memory for previous instances of the studied word.

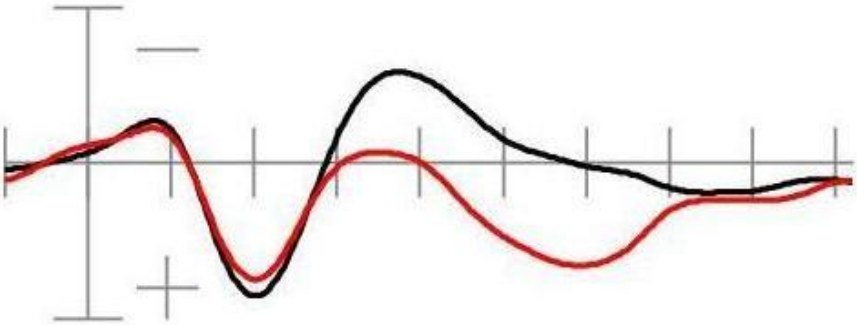


Figure 8. Repetition Effect for Study-Study Short Lag Items. Black Line = First Presentation, Red Line = Second Presentation.

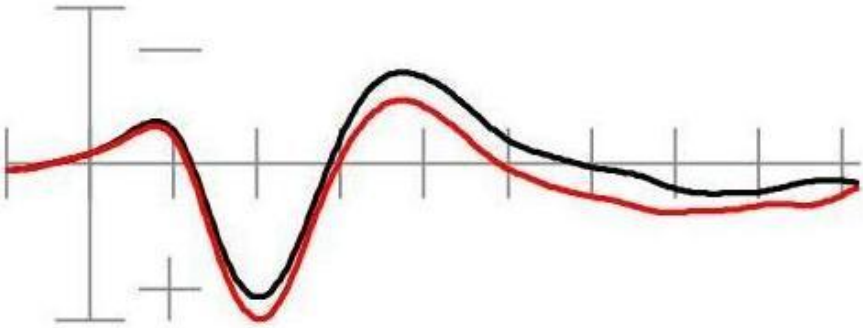


Figure 9. Repetition Effect for Study-Study Long Lag Items. Black Line = First Presentation, Red Line = Second Presentation.

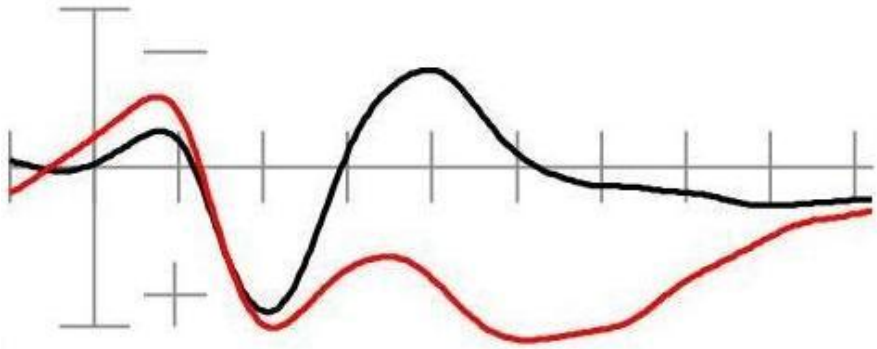


Figure 10. Repetition Effect for Study-Test Short Lag Items. Black Line = First Presentation, Red Line = Second Presentation.

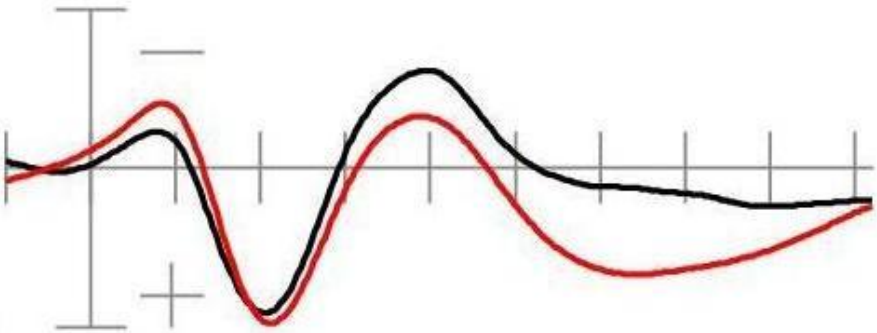


Figure 11. Repetition Effect for Study-Test Long Lag Items. Black Line = First Presentation, Red Line = Second Presentation.

3.2.2.3. Dm Effect

The traditional Dm effect was calculated by backsorting the ERPs elicited by the critical words based on subsequent memory for those words. For all conditions the Dm effect was largest over posterior channels, so analyses of the Dm effect were carried out using data from the nine electrodes showing the largest Dm effects across conditions: LMCe, MiCe, RMCe, LDPa, MiPa, RDPa, LMOc, MiOc, and RMOc. Figure 12 shows the backsorted ERPs for the once-presented condition. This comparison shows the typical Dm Effect, in which the subsequently remembered words elicit a broad positivity relative to the subsequently forgotten words.

For the first presentation of the study-study words, shown in Figure 13, the Dm Effect did not look like the typical Dm Effect that has been observed previously in the literature. It was not significant in the N400 or LPC time windows. The same pattern was apparent for the second presentation of study-study words at long lags, as shown in Figure 14.

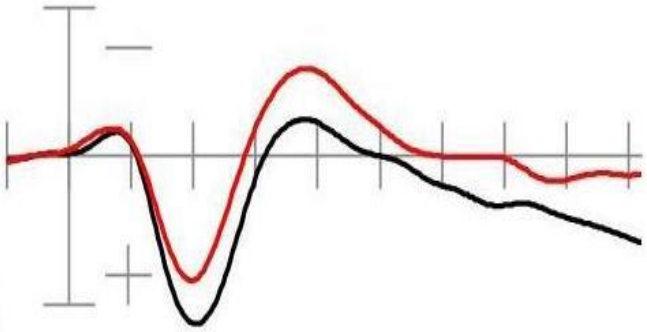


Figure 12. Dm Effect for Once-Studied Words. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

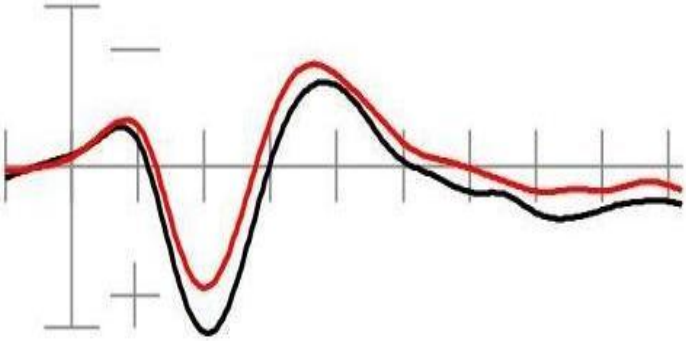


Figure 13. Dm Effect for Study-Study Items, First Presentation. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

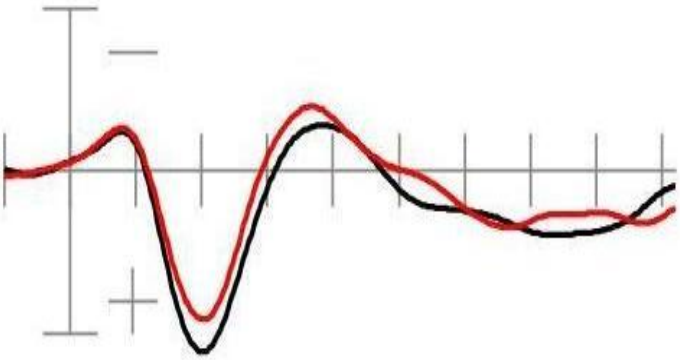


Figure 14. Dm Effect for Study-Study Second Presentation, Long Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

The Dm Effect for the second presentation of study-study words at short lag exhibited a different pattern, shown in Figure 15. The Dm Effect for those items exhibited the broad positivity for remembered items that is typically observed in Dm paradigms. However, the morphology of the effect was quite different than the standard Dm Effect, due to the smaller N400 and larger LPC for all words in this condition.

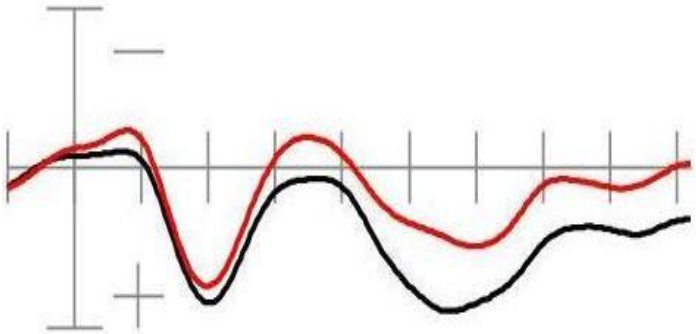


Figure 15. Dm Effect for Study-Study Second Presentation, Short Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

For the words that were studied and then tested, there was no Dm Effect whatsoever for the first presentation of the words, as shown in Figure 16.

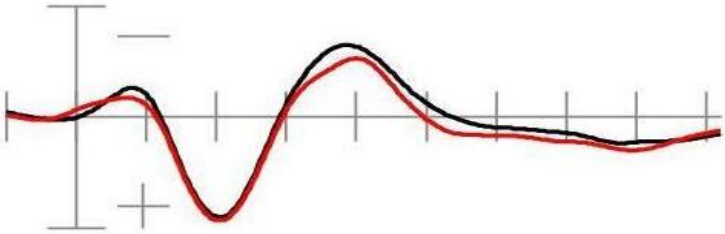


Figure 16. Dm Effect for Study-Test First Presentation. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

For the tested word, when it was presented at a short lag, the Dm Effect appeared once again, as shown in Figure 17. As with the study-study words, the morphology of the Dm Effect was influenced by the small N400 and large LPC of the tested words.

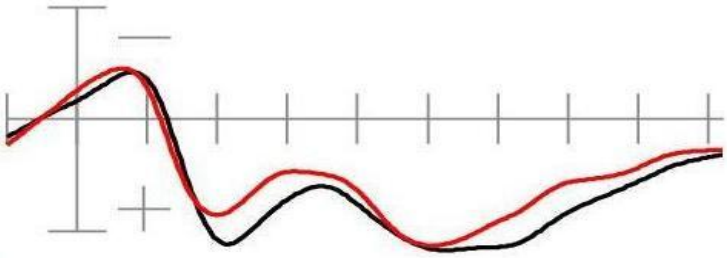


Figure 17. Dm Effect for Study-Test Second Presentation, Short Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

The second presentation of the study-test words in the long lag condition did not exhibit the typical Dm Effect. Instead, these items had a frontal negativity for subsequently remembered words, as shown in Figure 18.

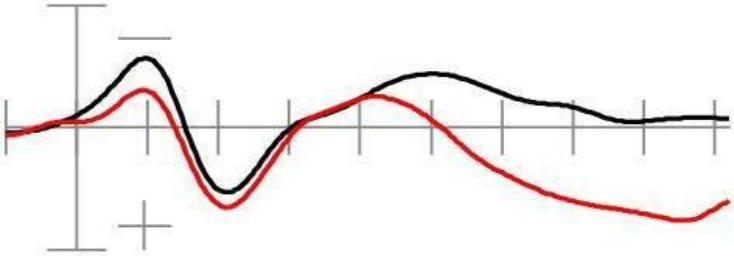


Figure 18. Dm Effect for Study-Test Second Presentation, Long Lag Items. Black Line = Subsequently Remembered Items, Red Line = Subsequently Forgotten Items.

3.2.3. Discussion

The results of Experiment 2 provide additional evidence in support of the findings from Experiment 1. The comparison of repetition effects across conditions revealed that the best behavioral memory performance was associated with a large LPC effect and little modulation of the N400. This pattern of brain activity suggests that the participants did not have semantic priming to help them recognize the words, as indexed by the small N400 repetition effect, but they were successfully able to recognize the words, as indexed by the large LPC. When recognizing the words during the study repetitions is easier, as indicated by a larger N400 repetition effect, participants gained less of a benefit to subsequent memory. These findings provide evidence for memory theories that suggest that "desirable difficulties" lead to the best learning and memory performance (Bjork, 1994). When the ERPs at study indicate that participants were expending more effort to retrieve information from memory, the participants had higher average performance on the subsequent memory test.

The results of Experiment 2 also replicate the findings of Experiment 1 with respect to the Dm and Dm-R effects. As in Experiment 1, there was no significant Dm effect for the first presentations of repeated or quizzed words. However, there were Dm-R effects that emerged when the second presentations of the study words were backsorted based on subsequent memory performance. These results suggest that subsequent memory performance is driven by the processing of the second presentation of the words, while the processing of the first presentation is not predictive of subsequent memory performance. In other words, if a studied word is presented more than once, the first presentation of that word is not equivalent to a study word that is presented only once. The repetitions of a word that intervene between its first presentation and a later memory test play a major role in whether or not that item is remembered. The opportunity for retrieval practice provided by the second presentation of a repeated item could be having a major influence on whether or not that item is subsequently remembered.

To test this hypothesis, we developed a computational model that can use ERPs recorded at study to predict whether or not a participant is likely to perform well on a subsequent memory test. The results of Experiments 1 and 2 point to retrieval practice as an important component of memory performance. The model was designed to classify participants as high performers or low performers based on how consistent their brain activity at study was with a retrieval practice strategy. The model is described in Section 4.

4. COMPUTATIONAL MODELING OF ERPS

The data collected in Experiments 1 and 2 was used to test different methods for constructing computational models of ERPs and ultimately to develop a model that can use classifications of ERPs recorded during study to predict participants' performance at test. For details about the various modeling approaches that were tested, see Appendix A. In this section, we outline the final modeling approach and how it was applied to the data from Experiments 1 and 2.

4.1. Computational Modeling of Experiment 1

The results of Experiment 1 showed that the best memory performance at test was associated with study items that were repeated at long lags during study. The ERPs elicited by these repetitions showed both a large N400 and a large LPC, indicating that the processing of the repetition was not facilitated by the prior presentation of that word, and that the participants were explicitly retrieving (or trying to retrieve) the prior presentation of that word from memory. As we argue above, this retrieval practice in the absence of facilitation provides a benefit to subsequent memory and is likely to play a role in the spacing effect in memory.

To further investigate the role of retrieval practice in memory, we developed a computational model to examine individual differences in the use of retrieval practice during this task. We hypothesized that some participants would make better use of the repeated words than others and that this difference would be reflected both in the ERPs elicited by the repeated words at study and in the participants' behavioral performance at test. If some participants made more of an effort to retrieve previous presentations of a repeated word, they should have a larger LPC and better memory performance at test. On the other hand, if some participants studied all of the words passively, rather than making an effort to recollect previous presentations of the words, those participants should have smaller LPCs on average and lower memory performance at test.

In order to test this hypothesis, we constructed a computational model that classified participants into "retrieval" and "passive study" groups based on the ERPs they produced in response to the words that were repeated after a long lag. This classification was achieved by constructing a naive Bayes classifier trained on two "known" conditions and applying this classifier to the "unknown" ERPs from the long-lag repetition condition. The average performance of the two groups of participants created by the classifier was then compared to test for differences in memory performance at test.

This process was competed for two different sets of known conditions. In both sets, the ERPs elicited by the words that were presented only once during the study session (with no associated repetition, synonym, or rhyme) were used as the known "passive study" condition. In one set, the known "retrieval" condition consisted of the ERPs elicited by the long-lag repeated items at test. This condition had the same words that were included in the unknown condition, but the ERPs were recorded during the memory test when the participants were explicitly asked to retrieve those words from memory. In the second comparison, the known "retrieval" condition consisted of the ERPs elicited by the new words that were presented during the memory test. In that case, the participants should attempt to retrieve the words but they will have no memory trace for those words from the study session. By modeling both successful and unsuccessful attempts at

retrieval, our goal was to determine the influence of retrieval success on any resulting differences in memory performance.

A significant challenge faced when constructing computational models from EEG signals is the low signal-to-noise ratio due to the presence of simultaneously recorded brain activity that is unrelated to the event of interest. This is often addressed by averaging all single-trial EEG recordings to form a grand average ERP. However, this approach removes most of the trial-to-trial variability and can result in the formation of a classifier that is not robust to variances present in the ERPs from the "unknown" condition.

To overcome these obstacles, we developed an approach that better balances variability and signal averaging. Our approach combined ensembling classification results from multiple models and randomized signal averaging of individual trial ERPs. Randomized signal averaging was accomplished using an n-choose-k approach to create a new set of ERPs for use in the classifier training step. We examined maximized signal averaging by using k=29 to select and average single trial EEG recordings in a time window from 100 ms pre-stimulus onset to 900 ms post-stimulus onset in order to create 40 ERP samples for each of the known conditions. For the passive study condition (the once-presented words), there were 80 trials per participant from which to choose. For the first retrieval condition (the long-lag repeated words at test) there were 40 single-trial ERPs and for the second retrieval condition (the new words presented at test) there were 320 single-trial ERPs from which to choose. The ERPs resulting from the n-choose-k averaging were then transformed via principal component analysis and the scores of the first five principal components were used as an uncorrelated feature set to train a naive Bayes classifier. The classifier was implemented by using MATLAB's classify function provided in the Statistical Toolbox with the "diaglinear" discriminant function. This process was then repeated 50 times using a new random seed to randomize the single-trial ERPs chosen for signal averaging from the n-choose-k trial selection process. In this way, each model was exposed to different signal averaging in the unknown condition ERP samples while maintaining a balanced number of training examples across the known conditions.

4.1.1. Modeling Results

The ERPs from 29 of the participants was classified as described above. One participant was excluded from the classification due to an insufficient number of trials in the critical conditions. The performance of the classifier was estimated using sample-out cross validation. The mean area under the receiver-operator curve (AUC) for sample-out cross validation over all models and all subjects was 0.99. The standard deviation of the mean sample-out cross validation AUC for each subject was 0.01. These cross validation results provide confidence that the feature extraction and classification methods are well suited to model the ERPs.

For classification of the unknown ERPs, a full model was constructed with all samples from each of the 50 randomly constructed training sets described above. This model was then used to classify the unknown ERPs as belonging to the passive study or retrieval groups. If more than half of the models for a given participant classified that person’s unknown ERPs as retrieval ERPs, that participant was placed in the retrieval group. Conversely, if more than half of the

models for a participant classified that person's unknown ERPs as passive study ERPs, that participant was placed in the passive study group.

For the first comparison, in which the known retrieval condition was defined as the long-lag repeated items presented at test, 12 of the 29 participants were placed in the retrieval group and 17 were placed in the passive study group. The two groups' memory performance at test was compared for the words in the unknown condition, the long-lag repeated words. The participants in the retrieval group were 78% correct on average while the participants in the passive study group were 69% correct. The difference in performance between the two groups was significant ($t(21) = 2.71$, $p < 0.01$). In fact, the participants in the retrieval group outperformed the participants in the passive study group across almost all conditions, as shown in Figure 19.

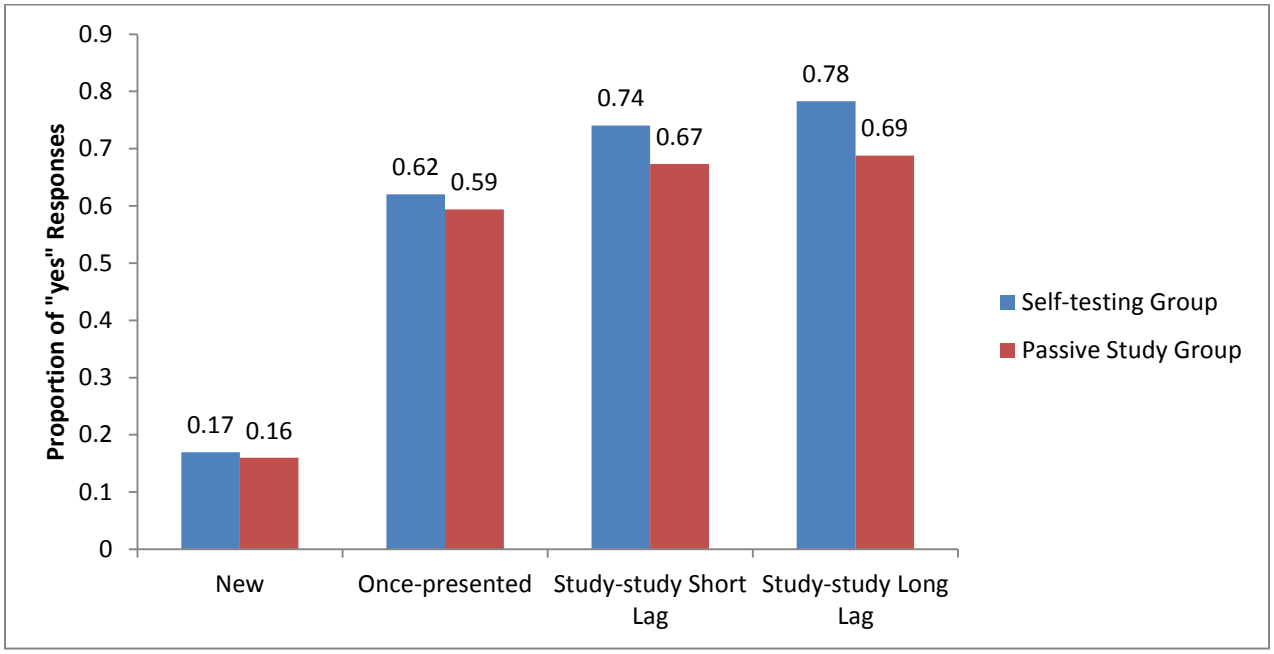


Figure 19. Proportion of "yes" responses for each test condition for the subject groupings created by the computational model.

For the second comparison, in which the known retrieval condition was defined as the new items presented at test, 15 of the 29 participants were placed in the retrieval group and 14 were placed in the passive study group. Once again, the participants classified in the retrieval group had significantly better memory for the long-lag repeated words than the participants in the passive study group (76% correct versus 69% correct, $t(25) = 2.02$, $p < 0.03$). Once again, the participants in the retrieval group outperformed the participants in the passive study group in most conditions, as shown in Figure 20.

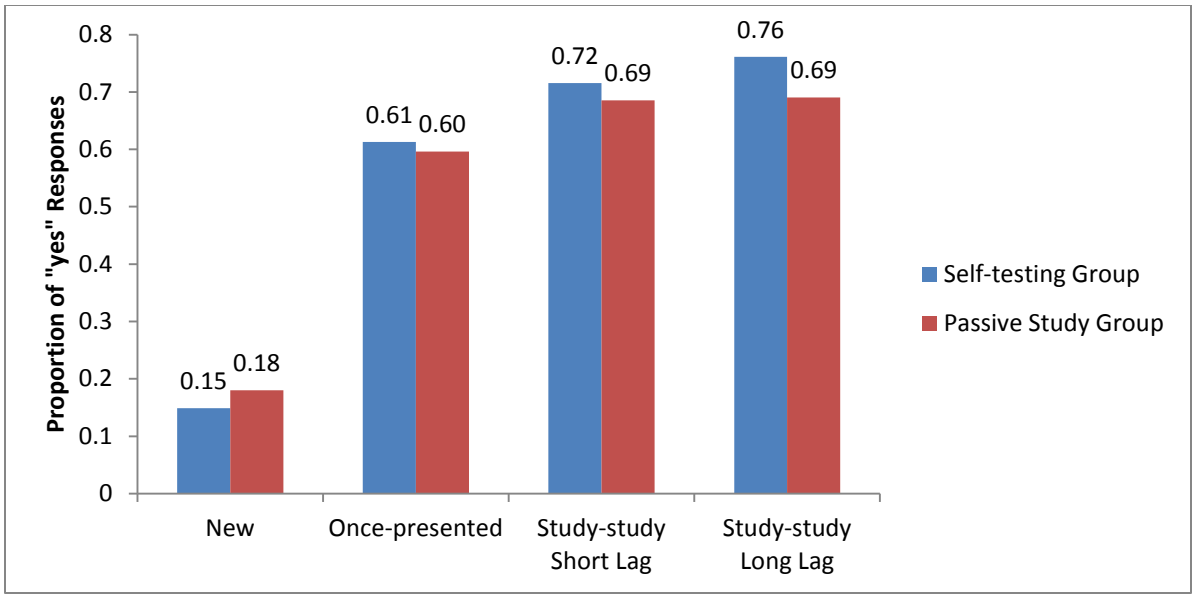


Figure 20. Proportion of "yes" responses for each test condition for the subject groupings created by the computational model.

4.2. Computational Modeling of Experiment 2

The modeling results from Experiment 1 demonstrated that ERPs recorded at study can be used to predict a participants' level of memory performance at test. We attempted to replicate this result by applying the same model to the data from Experiment 2.

For the modeling of the Experiment 2 ERPs, the known conditions were the first presentations of studied words (passive study condition) and the words that were quizzed after a long lag (retrieval condition). In the passive study condition there were 278 single trial EEG recordings from which to choose and for the retrieval condition there were 40 single trial EEG recordings. As in the Experiment 1 modeling, randomized signal averaging of individual trial ERPs was accomplished using an n-choose-k approach to create a new set of ERPs for use in the classifier training step. We examined maximized signal averaging by using k=39 to select and average single trial EEG recordings in a time window from 100 ms pre-stimulus to 900 ms post-stimulus to create 40 ERP samples for each of the two known conditions. The resulting ERP samples were then transformed via principal component analysis and the scores of the first five principal components were used as an uncorrelated feature set to train a naive Bayes classifier. This process was then repeated 50 times using a new random seed to randomize the single trial EEG recordings chosen for signal averaging from the n-choose-k trial selection process. In this way, each model was exposed to different signal averaging in the unknown condition ERP samples while maintaining a balanced number of training examples across the two known conditions.

4.2.1 Modeling Results

The ERPs from 23 of the 24 participants was classified as described above. One participant was excluded from the classification due to an insufficient number of trials in the critical conditions. The performance of the classifier was estimated using sample-out cross validation. As in

Experiment 1, the mean area under the receiver-operator curve (AUC) for sample-out cross validation over all models and all subjects was 0.99. The standard deviation of the mean sample-out cross validation AUC for each subject was 0.01.

For classification of the unknown ERPs, a full model was constructed with all samples from each of the 50 randomly constructed training sets. This model was then used to classify the unknown ERPs as belonging to the passive study or retrieval groups. Examination of the number of models classifying the unknown ERPs as belonging to the study group identified eighteen subjects whose brain activity was consistent with their previously used study strategy. For this group of eighteen subjects, more than 97% of the models for each subject identified the unknown ERPs as belonging to the passive study class. Another group of five of the subjects exhibited brain activity that was consistent with the retrieval practice elicited by the quizzed words. The number of models identifying the unknown ERPs as belonging to the test class varied with subject and ranged from 22% to 80% of the 50 models constructed for each subject. This variation is indicative of individual differences and may indicate that the retrieval practice strategy was employed with different frequency by each subject.

To test the model's classification performance, we compared the behavioral memory performance across the two groups of participants. As predicted, the participants whose brain activity in the unknown condition was consistent with their brain activity in the retrieval practice condition had better memory for the twice-studied items than participants whose brain activity was consistent with passive study. On average, the participants in the former group correctly recognized 28.6 out of 40 words (71.5%) from the twice-studied condition, while the participants in the latter group correctly recognized 18.1 out of 40 words (45.2%). Welch's t-test showed that the performance of the two groups was significantly different [$t(9.4) = 3.82$, $p < 0.01$]. Figure 21 shows the grand average ERPs for the unknown condition and the test condition for the two groups. As in the models of Experiment 1, the participants assigned to the retrieval practice condition by the model outperformed the other participants across all of the test conditions, as shown in Figure 22.

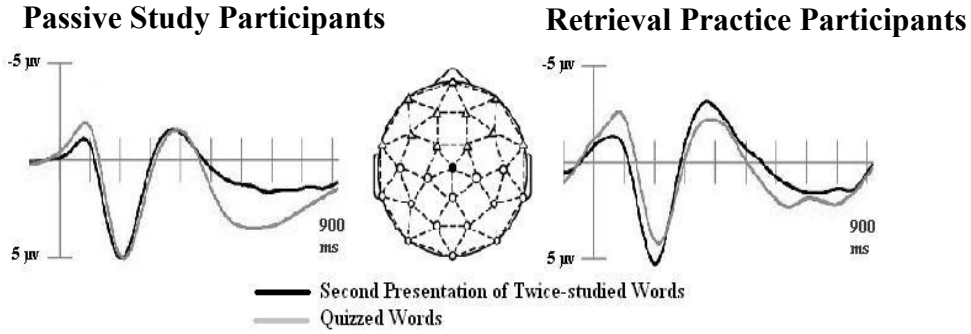


Fig. 21. Grand average ERPs to the unknown condition, the *second presentation of twice-studied words (black line)*, and the test condition, the *quizzed words (gray line)*. ERPs are shown at the midline central (MiCe) electrode. The participants whose brain activity was consistent with passive study in the unknown condition are shown on the left and the participants whose brain activity was consistent with a retrieval practice strategy are shown on the right.

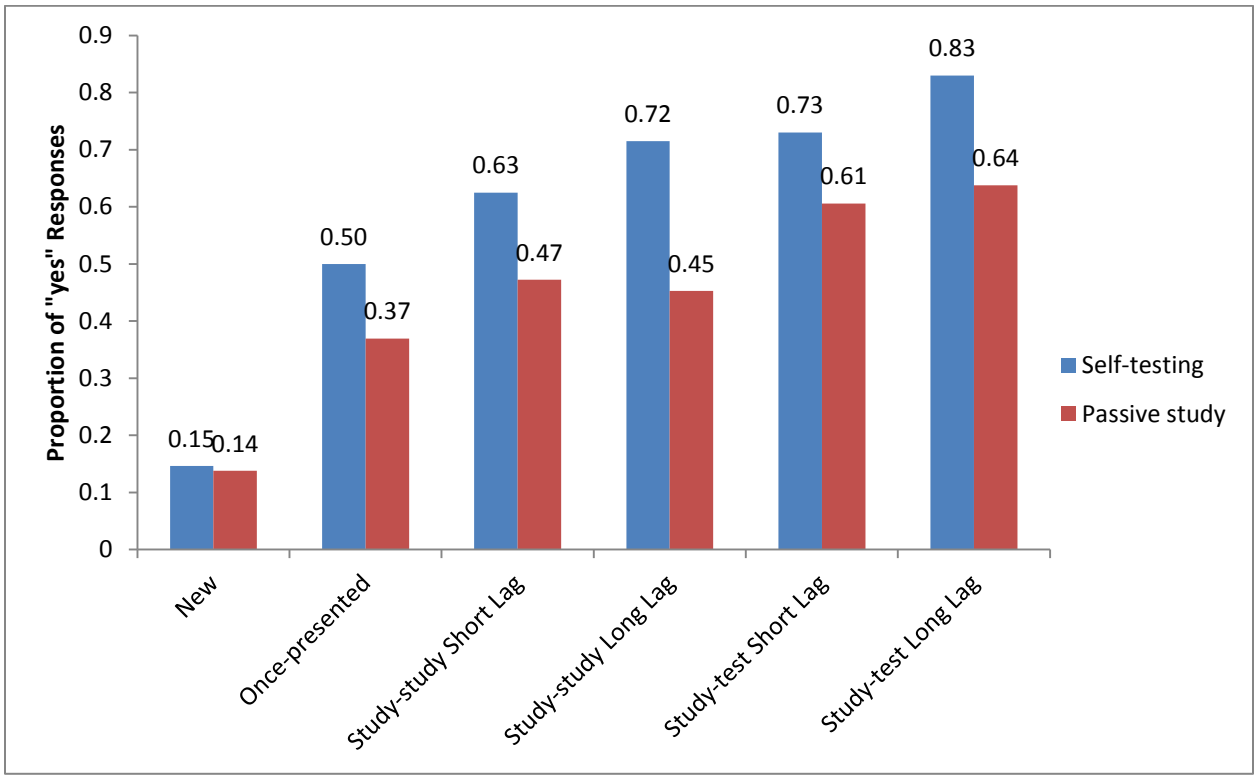


Figure 22. Proportion of "yes" responses for each test condition for the subject groupings created by the computational model.

4.3. Discussion

The modeling results from Experiments 1 and 2 indicate that ERPs elicited under known conditions can be modeled and used to classify ERPs from an unknown condition. In these experiments, the known conditions included passive study conditions and conditions in which participants were tested or quizzed on previously studied words, leading the participants to engage in retrieval practice. The unknown condition was the second presentation of repeated study items. For those items, participants might retrieve the first presentation of the word from memory, adopting a retrieval practice strategy on their own. Previous research on study strategies and cognitive biases led us to predict that few participants would spontaneously engage in retrieval practice, but those that did would outperform the other participants for the words in that condition.

As we predicted, the average memory performance across all participants was lower for the words that were studied twice than for the words that were studied and then quizzed. Using the model, we identified groups of participants whose brain activity was consistent with use of a retrieval practice strategy. Those subsets of participants performed significantly better than the other participants on the subsequent memory test in both experiments.

These experiments and the development of a model that is predictive of later performance represent crucial steps toward using recorded brain activity to improve human memory

performance. We have identified patterns of brain activity that are associated with the use of an effective memory strategy and developed a model that can predict which participants are using that strategy and which are not. This approach has not previously been attempted with ERP data, so Experiments 1 and 2 provide a demonstration and replication of an important new finding.

In Experiment 3, we expanded upon this research by conducting a large-scale memory training study. The results of Experiments 1 and 2 demonstrate that using an effective memory strategy, such as retrieval practice, has a significant impact on performance at test and is reflected in brain activity at study. In Experiment 3, we gave participants a variety of memory tasks before and after memory training sessions. In the training sessions, they practiced either a memory strategy (mental imagery) or practiced non-strategic tasks intended to increase their basic working memory capacity. We investigated the effects of strategic and non-strategic training on both memory performance and on brain activity.

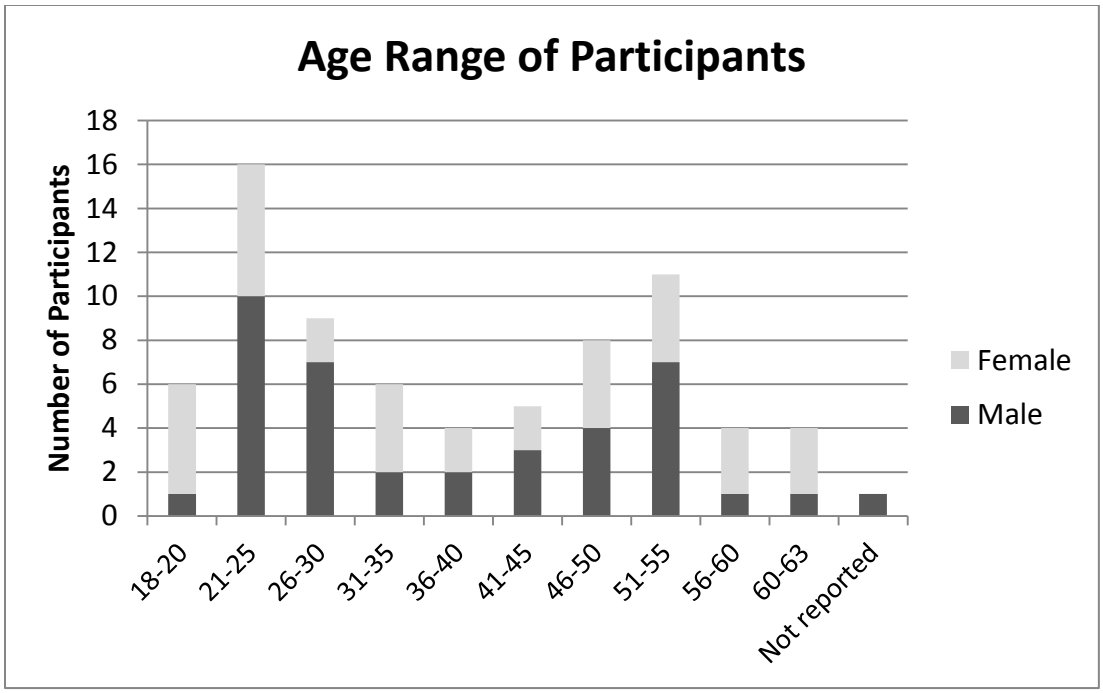
5: EXPERIMENT 3 - THE EFFECTS OF COGNITIVE TRAINING ON MEMORY

5.1. Method

Each participant in the experiment completed tasks over the course of a five-week period. During the first week, participants completed six baseline tasks and their brain activity was recorded using electroencephalography (EEG) during three of the six tasks. During the next three weeks, participants completed memory training sessions that differed based on the training group to which they were assigned. Participants assigned to the mental imagery training group completed three training sessions (one per week) and participants in the working memory training group completed 14 training sessions (4-5 per week) during the three week training period. Participants assigned to the control group did not complete any tasks during the training period. At the end of the training period, all participants completed the same six baseline tasks for a second time. Each of the baseline and training tasks is described in detail below.

5.1.1. Participants

Eighty-six participants recruited from the employee population of Sandia National Laboratories participated in this experiment and were paid for their time. All were right-handed, had no early exposure to languages other than English, and had no history of neurological disease or defect. Participants were assigned semi-randomly to one of the three training groups (efforts were made to balance the three groups based on age and gender). Eight participants dropped out of the study before completing all of the sessions and four additional participants failed to follow instructions and were excluded from the data analysis. Of the remaining 74 participants, 25 (12 female) were in the control group, 24 (10 female) were in the imagery training group, and 25 (13 female) were in the working memory training group. The mean age for all of the participants was 37 (range 18-63). The mean ages for each group were 37 for the control group (range 18-61), 39 for the imagery training group (range 18-63) and 35 for the working memory training group (range 20-63). As the demographics of the Sandia employee population are quite different from college student populations, the distributions of age and educational background for the participants in this study are shown in Figure 23.



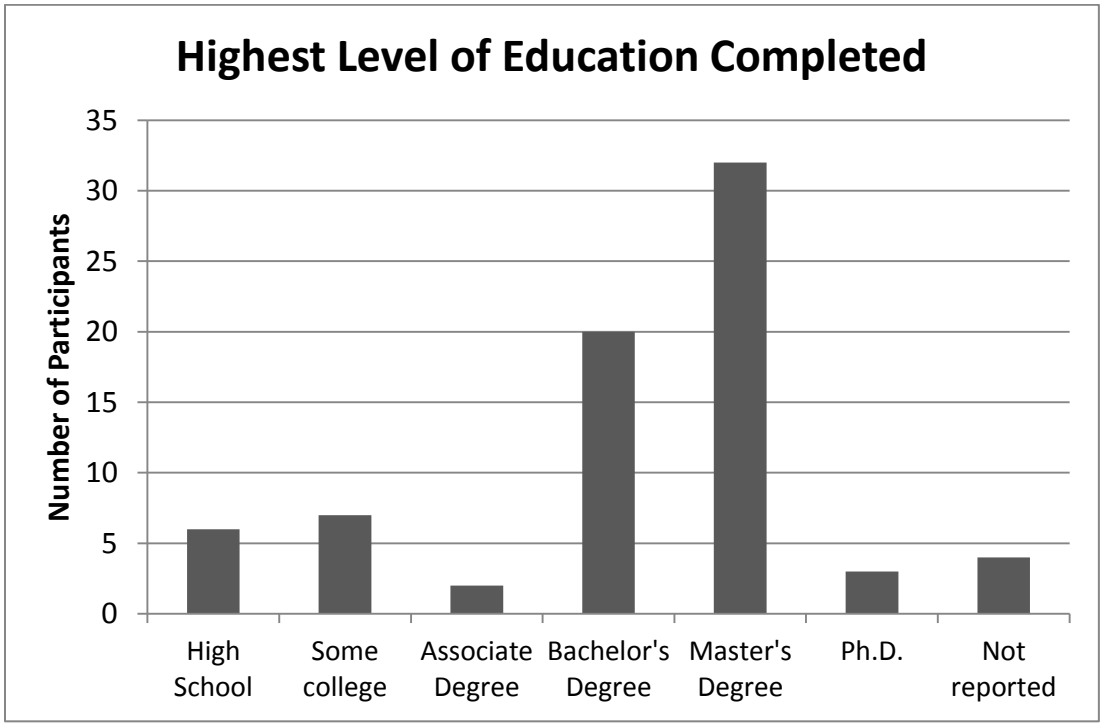


Figure 23. Distributions of age and educational background for the participants in the memory training study.

5.2. EEG Baseline Tasks

During the pre- and post-training baseline sessions, participants completed three memory tasks while their brain activity was recorded using EEG. The first task was a test of recognition memory, the second task was designed to induce different types of recognition memory errors, and the third task was a test of associative memory.

5.2.1. General Procedure

While completing all three EEG baseline tasks, participants were seated in a dimly-lit, sound-attenuating booth where they viewed the stimuli on a computer monitor. Throughout each task, a fixation cross was shown in the center of the screen and the participants were instructed to keep their eyes on the fixation cross at all times during the task. They were asked to avoid blinking while a word or picture was on the screen. When the tasks required a response from the participants, they pressed one of two buttons on a game controller, labeled “yes” and “no.” The assignment of the response buttons was counterbalanced across participants so that half of the participants used their right hand to press the “yes” button and half of the participants used their left hand. During the pre-training EEG session, participants completed a practice session to familiarize them with each task.

5.2.2. Procedure for EEG recording

The EEG was recorded from 128 silver/silver-chloride electrodes embedded in an elastic cap (ANT WaveGaurd, “Duke” layout), referenced to an average of all electrodes. All of the electrodes were tested before recording to ensure that their impedance was below 50 KOhms. The EEG was recorded using a 128-channel, high-impedance ANT WaveGuard amplifier with active shielding and was digitized with a sampling rate of 1024 Hz.

Event-related potentials (ERPs) were computed at each electrode for each experimental condition by averaging the EEG data from 100 ms before the onset of a stimulus until 920 ms after stimulus onset. Trials containing blinks were corrected using EEGLAB’s (Delorme & Makeig, 2004) independent components analysis to identify and remove activity related to blinks. Trials containing other artifacts such as muscle activity of signal drift were excluded from the averages. The mean amplitude of the ERPs within time windows of interest was calculated using data digitally filtered off-line using a bandpass filter of 0.2 to 20 Hz.

5.2.3. Recognition Memory Task

In the recognition memory task, participants were shown a list of common English nouns and were asked to memorize them for a subsequent recognition test. Some of the nouns were repeated at short (one intervening item) or long (nine intervening items) lags during the study blocks, while other study items were quizzed at short or long lags during the study blocks. All of the studied items were subsequently tested in a recognition memory test, intermixed with an equal number of new, unstudied items.

5.2.3.1. Design. The critical variables in the recognition memory task were the conditions in which the words were studied (studied once, studied twice, or studied and then quizzed during the study block) and the lag between the study-study and study-quiz repetitions (one or nine intervening items). Both of these variables were manipulated within-subjects. The dependent variable for the behavioral test was yes/no recognition for the studied words in each condition and at each lag.

5.2.3.2. Materials. The recognition memory task used a list of 1344 words, all of which were common English nouns. The average length of the nouns was 5 letters and their average frequency was 55.67 (based on the Kucera and Francis [1967] norms included in Balota et al., 2002). The words were assigned to counterbalanced experimental lists such that every word appeared in every study and test condition across lists.

The experimental lists were divided into six study-test blocks with equal numbers of each item type in each block. The words were placed in a pseudorandom order within the blocks such that no more than three items in the same condition appeared in sequence. Within each study-test block, there were 28 words that were studied once, 14 words that were studied twice with a short lag between repetitions, 14 words that were studied twice with a long lag between repetitions, 14 words that were studied and then quizzed after a short lag, and 14 words that were studied and then quizzed after a long lag. In addition to the studied items, there were 28 words that served as new items for the quizzes within the study blocks (these words were quizzed but had not been studied) and 84 words that served as new, unstudied items in the subsequent recognition test. In total, each study block contained 112 study words (including repeated study words) and 56 quizzed words. Each test block contained 168 test words, half of which had been studied and half of which were new.

Three of the study-test blocks were presented to each participant during the pre-training EEG session and the other three were presented during the post-training EEG session. The placement of the blocks (pre- or post-training) was counterbalanced across participants.

5.2.3.3. Procedure

The participants were instructed that they would be tested on their memory for a list of study words. All of the words were presented immediately above the fixation cross in white 48-point Arial font on a black background. Within the study blocks, each study word was preceded by a dot that was presented above the fixation cross for one second. For the study words, the dot was yellow and for the quiz words the dot was red. Participants were instructed that the yellow dot indicated an upcoming study word and that they should silently read that word and try to remember it for later. They were told that the red dot indicated that the next word was a quiz word, and that following the word they should press a button to indicate whether or not they had studied that word earlier in the session. The study or quiz word was presented 600-800 ms after the dot disappeared and remained on the screen for one second. After each quiz word, a red question mark appeared on the screen and remained until the participant pressed the “yes” or “no” response button.

At the end of each study block, participants took a short break before beginning the test block. All of the words in the test block were presented in the same way as the quizzed items from the study block. Each test word was preceded by a red dot and followed by a question mark. While the question mark was on the screen, participants pressed the "yes" or "no" button to indicate whether or not they remembered studying that word during the study block. It took participants approximately 10 minutes to complete each study block and 12 minutes to complete the corresponding test block.

5.2.4. Memory Error Task.

In the memory error task, participants studied adjective-noun pairs (such as "cozy couch" and "cluttered room") and were asked to memorize them for a subsequent memory test. During the test, participants were tested on old pairs, semantic lure pairs, conjunction lure pairs, and new pairs. In the semantic lure pairs, the originally studied noun was replaced with a synonym (such as "cozy sofa"). In the conjunction lure pairs, two studied pairs were recombined so that the tested pair contained the adjective from one studied pair and a noun from a different pair (such as "cozy room"). In the new pairs, neither word had been studied previously.

5.2.4.1. Design.

The critical variable in the memory error task was the condition in which the word pairs were tested (old pair, semantic lure, conjunction lure, new pair). This variable was manipulated within-subjects and the dependent variable for the behavioral test was yes/no recognition performance for each test condition.

5.2.4.2. Materials.

The 780 adjective-noun pairs used in the memory error task were comprised of 520 adjectives and 600 nouns. The items were created and counterbalanced so that each adjective-noun pair appeared equally often as a study pair and as a test pair. In addition, each item was studied and tested in every condition (as an old parent/old test item, as a conjunction lure parent/conjunction lure, and as a semantic lure parent/semantic lure). Eighty adjective-noun pairs were used exclusively as new, unstudied pairs at test. The words used in the new pairs were matched to the studied items in terms of length and frequency.

The counterbalancing produced 12 experimental lists, each of which contained 80 pairs that appeared in the same form at study and test (old pairs), 80 pairs in which the studied noun was replaced with a synonym at test (semantic lure pairs), 160 pairs that served as the parents of conjunction lures at study (giving rise to 80 conjunction lures at test), and 80 new pairs that were presented only at test. The experimental lists were divided in half so that half of the items were studied and tested during the pre-training EEG session and the remaining items were studied and tested during the post-training EEG session. Thus, during each session, the participants studied a total of 160 adjective-noun pairs and were tested on 160 adjective-noun pairs (40 old pairs, 40 semantic lures, 40 conjunction lures, and 40 new pairs).

5.2.4.3. Procedure.

The participants were instructed that they would be tested on their memory for adjective-noun pairs. The words were presented one at a time, immediately above the fixation cross in white 48-point Arial font on a black background. During the study session, each word pair was preceded by a yellow dot that was presented above the fixation cross for one second. After the dot disappeared, there was a variable interstimulus interval (ISI) of 600-800 ms and then the adjective was presented for 1000 ms. The adjective was followed by a blank screen that was presented for 500 ms and then the noun was presented for 1000 ms. During the test phase, the pairs were presented in the same way except that each test pair was preceded by a red dot and followed by a question mark. The participants were instructed to press the "yes" or "no" button while the question mark was on the screen to indicate whether or not they remembered studying that word pair. They were warned that some of the test pairs would be similar to, but slightly different from, the studied pairs. They were instructed that they should only press the "yes" button if they were sure that the word pair was exactly the same as one they had studied. It took participants approximately 12 minutes to complete the study block and 15 minutes to complete the test block.

5.2.5. Associative Memory Task.

In the associative memory task, participants studied pairs of unrelated pictures and were asked to memorize them for a subsequent memory test. During the test, participants were presented with intact, rearranged, and new picture pairs. For each pair, participants indicated whether or not they remembered studying the pictures and whether or not they remembered studying the pictures together.

5.2.5.1. Design.

The critical variable in the associative memory task was the condition in which the picture pairs were tested (old pair, rearranged pair, or new pair). This variable was manipulated within subjects and the dependent variable for the behavioral test was yes/no recognition performance for each condition.

5.2.5.2. Materials.

The stimuli used in the associative memory study consisted of 480 simple line drawings that were paired so that the two items in the pair were not related in any way (reference from Eric). The picture pairs were divided into eight counterbalanced lists such that all of the pairs appeared in every study and test condition across lists. In each list there were 60 picture pairs that were studied and then tested as intact pairs, 120 picture pairs that were studied and then rearranged at test to form 60 rearranged pairs (with the two tested pictures taken from two different studied pairs), and 60 picture pairs that were presented as new, unstudied pairs at test. The experimental lists were divided in half so that half of the items were studied and tested during the pre-training EEG session and the remaining items were studied and tested during the post-training EEG session. Thus, during each session, the participants studied a total of 90 picture pairs and were tested on 30 intact pairs, 30 rearranged pairs, and 30 new pairs.

5.2.5.3. Procedure.

The participants were instructed that they would be tested on their memory for picture pairs. As in the word pair task, each pair was preceded by a dot that was presented above the fixation cross for one second. The dot was yellow during the study phase and red during the test phase. After the dot disappeared, there was variable ISI of 600-800 ms before the first picture in the pair was presented. The pictures were centered over the location of the fixation cross and were presented for one second each, with a 500 ms blank screen presented between the two pictures in the pair. In the test phase of the task, the second picture in the pair was followed by two questions. The first question read "Did you see these pictures before?" and the second question read "Were these pictures paired together before?" Each question remained on the screen until the participant pressed the "yes" or "no" button. Participants were instructed that they would not see pairs in which one picture had been presented and the other had not. It took participants approximately 8 minutes to complete the study block and 10 minutes to complete the test block.

5.3. Behavioral Baseline Tasks

The three pre- and post-training baseline tasks that were conducted without EEG recordings included the Deese-Roediger-McDermott (DRM) Task and two measures of working memory span, the Rotation Span Task and the Listening Span Task. Participants had the option of completing the three behavioral tasks in the lab or by connecting to a lab computer via a remote desktop connection from their own computer.

5.3.1. Deese-Roediger-McDermott (DRM) Task.

In the DRM Task (Deese, 1959; Roediger & McDermott, 1995), participants were presented with lists of related words (e.g. "hot, winter, ice, snow, freeze") and then were tested on old words (such as "hot"), new unrelated words (such as "dog"), and lure words that were related to the studied lists but had not been studied (such as "cold").

5.3.1.1. Design.

The critical variable in the DRM task was the test condition of the words (old, lure, or new) and the dependent variable was yes/no recognition performance for each condition. In addition, participants were asked to make a remember/know judgment for each word to which they responded "yes" on the recognition test.

5.3.1.2. Materials.

The thirty-six DRM lists used in this experiment were drawn from Roediger, Watson, McDermott and Gallo (2001). Each list contained 15 study words that were related to a theme word. Half of the lists were presented during the pre-training session and half were presented during the post-training session[2].

[2] The theme words for the lists used in the pre-training session were: anger, car, cold, command, cup, doctor, fruit, high, justice, lamp, lion, man, mountain, music, river, rough, smell, and soft. The theme words for the lists

5.3.1.3. Procedure.

The participants were instructed that they would see several lists of words and would be tested on their memory for the words. The 18 DRM lists that were studied in each session were presented in a random order. The words in the lists were presented one at a time, in black 40-point font. Each list was preceded by a screen reading "Next List" that was presented for two seconds. Then each word in the DRM list was presented for one second. Following the presentation of all 18 study lists, the participants completed a short distractor task in which they were asked to answer a series of 40 simple math problems as quickly and accurately as possible.

The memory test began once participants had finished all of the problems in the math distractor task. Participants were instructed that they would see a series of test words and that they should press the "y" key on the keyboard if they remembered studying the word and the "n" key if they did not remember studying the word. They were then instructed that each time they pressed the "y" key, they would be asked if their memory for the word was a "Type A" or "Type B" memory. A "Type A" memory was defined as a memory where participants could recall the exact thoughts that came to mind at the time they studied that word, such as personal associations, mental imagery, emotions, or other details related to studying that word. A "Type B" memory was defined as cases where the participant "just knows" that he or she studied the word, but could not recall any specific associations or details related to studying the word. During the memory test, the participants were presented with 108 words, including 18 lures (the theme words of the 18 DRM lists), 54 old words, and 36 new words that were not related to any of the studied words. Each test word remained on the screen until the participant pressed the "y" or "n" key. Each time the participants pressed the "y" key, they were asked to indicate what type of memory they had by pressing the "a" key for a Type A memory or the "b" key for a Type B memory. The definitions of the Type A and Type B memories were shown at the bottom of the screen to remind participants of the meaning of the terms.

5.3.2 Listening Span Task.

Based on Daneman & Blennerhassett (1984), the listening span task required participants to recall a sequence of symbols in the order in which they were presented. The presentation of the symbols was interleaved with the auditory presentation of sentences. Participants had to indicate whether the sentences made sense or not. Participants practiced the two tasks separately, and then performed both in the dual task phase.

5.3.2.1. Design.

The dependent variable in listening span is the total number of symbols recalled in the correct order during a dual-task test phase. Participants had to maintain a high level of performance in the sentence judgment task, (85% accuracy) in order for their data to be included in the analysis.

---

used in the post-training session were: bread, chair, city, flag, foot, girl, needle, pen, rubber, shirt, sleep, slow, smoke, spider, sweet, thief, trash, and window.

5.3.2.2. Materials.

Materials for the memory task were nine black Wingdings symbols in size 24 font presented against a white background:

🔔 ☎ 🕯 ☺ ✂ 💣 ⚑ ✎ ☠

The secondary task was comprised of 110 spoken sentences, half of which were sensible ("They gave the waiter a tip even though he was rude"), and half of which were not ("The children were summer and wanted their parents to come home"). The same speaker was used for all sentence recordings.

5.3.2.3. Procedure.

The listening span task had three phases, the first of which was a memory task involving the sequences of symbols. Participants saw symbols presented for 1000 ms in the center of a computer screen. For recall, a 3x3 grid appeared on the screen containing all possible symbols. Three additional buttons appeared along the bottom of the grid labeled "next," "clear," and "blank." Participants responded by clicking on the grid area corresponding to the symbols from the previous sequence in the order they were presented. Symbols appeared in a separate grid between the response grid and buttons as they were clicked, allowing participants to evaluate their response accuracy. If a participant made an error while clicking, they used the "clear" button to remove the most recent response. If they forgot a symbol in the sequence, participants were instructed to click the "blank" button. When the participants were satisfied with the sequence of symbols in their responses, they clicked "next." Following each response screen, participants saw feedback on their performance (i.e., "you recalled X of Y items correctly").

The second task required participants to judge the sensibility of sentences. Sentences were presented using headphones, which participants adjusted to a comfortable volume. During sentence presentation, participants received instructions to click the mouse once they could tell whether the sentence was made sense or not, at which point the sentence stopped playing and they responded by clicking either the yes or no button. Following this response, feedback on accuracy appeared on the screen.

During the performance phase, participants saw a new symbol after judging each sentence. The recall screen appeared after a sequence of 4 to 8 symbols had accrued. Each participant saw two sequences of 4 and 5 symbols, and three sequences of 6, 7, and 8 symbols. The different sequence lengths were randomly ordered for each participant.

5.3.3. Rotation Span Task.

Based on Shah & Miyake (1996), this task required participants to recall sequences of arrows of varying length and orientation. The presentation of the arrows was interleaved with the presentation of letter characters. Participants had to make a judgment as to whether the letters appeared normally or backwards. After each block, participants were asked to recall the sequence of arrows. As in listening span, participants practiced the two tasks separately, and then performed both in the dual task phase.

5.3.3.1. Design.

The dependent variable in rotation span is the total number of items (arrows) recalled during a dual-task test phase. Participants had to maintain a high level of performance in the rotation judgment task (85% accuracy) in order for their data to be included in the analysis.

5.3.3.2. Materials.

For the memory task, the item set was comprised of pictures of long and short arrows at nine orientations, (0°, 45°, 90°, 135°, 180°, 225°, 270°, 310°, 355°). The secondary task used five letters, (R, L, J, G, and F), at eight different orientations, (45°, 90°, 135°, 180°, 225°, 270°, 310°, 355°), with both normal and backwards versions (flipped along the vertical axis).

5.3.3.3. Procedure.

The first phase of rotation span involved practicing the recall task (recalling the sequences of arrows). Sequences of arrows that varied in length and orientation were presented in random order. Arrows were always white and presented against a black background for 1000 ms. After the presentation of a sequence of arrows, two 3x3 grids containing all of the possible arrow stimuli appeared on the screen. The center square of each grid was blank, and the surrounding 8 blocks contained arrows in all of the 8 possible orientations. Participants clicked on these arrows to indicate which arrows had appeared in the previous sequence, in the order that they appeared. As each arrow was clicked, it appeared in a blank space on the screen below the grids. The response screen also contained three buttons reading, “clear,” “blank,” and “next.” Clicking “clear” removed the last mouse response, in case the participants accidentally clicked on the wrong stimulus. Clicking “blank” allowed participants to note that they had forgotten which arrow appeared in a given position in the sequence without altering the order of the arrows they did remember. When the participants completed their responses for the current list, they clicked the “next” button to advance to the next trial.

The second task required participants to judge whether letters were either correctly oriented or backwards. The stimuli consisted of five letters, presented either normally or backwards, rotated to appear in eight different orientations. The letters were white, presented on a black background. In the dual task phase, participants saw a new arrow after responding to the orientation of each letter. The recall screen appeared after a sequence of two to five arrows had been presented. Each participant saw three sequences of each length in a random order.

5.4. Memory Training

5.4.1. Mental Imagery Training.

In the three weeks in between the pre-training and post-training baseline sessions, 24 of the participants completed three memory training sessions in which they practiced using a mental imagery strategy to remember word lists for a free recall test. The training sessions became more

difficult as the participants progressed by using longer word lists, shorter encoding times, and more words with low imagability.

5.4.1.1. Materials.

The memory tests used in the mental imagery training consisted of 168 nouns. Care was taken to ensure that none of the words used in the training sessions appeared in any of the pre-training or post-training baseline tasks. Of the 168 nouns, 49 had low imagability (ratings below 400 in the norms included in the MRC Psycholinguistic Database, Wilson, 1988) and the remainder had high imagability (ratings above 550).

5.4.1.2. Procedure.

For all of the mental imagery training sessions, participants had the option of completing the training in the lab or by logging in to a lab computer via a remote desktop connection from their own computer. The training consisted of three sessions and participants were asked to complete the sessions once a week for three weeks. Each session took approximately half an hour to complete.

Participants began the first training session by completing two questionnaires about their use of mental imagery, the Object-Spatial Imagery Questionnaire (OSIQ; Blajenkova, Kozhevnikov & Motes, 2006) and the Vividness of Visual Imagery Questionnaire (VVIQ; Marks, 1973). The participants were then given examples of mental imagery. The examples, which included both concrete and abstract concepts, explained that creating detailed and unusual mental images could be helpful for remembering information. After a short practice session in which participants were asked to generate and describe mental images for a short list of words, the training provided examples of grouping several mental images into one scene to increase their memorability. The participants were then asked to practice the mental imagery strategy by memorizing two lists of words, each of which was followed by a recall test. During the first practice list, the participants controlled the presentation of the study words and could view each word for as long as they wanted. For the second practice list, each word was presented for three seconds. Each practice list contained 10 words and participants had 10 chances to enter the words during the recall test. They received feedback after entering each word to indicate whether or not each entry was correct. After the memory test, the participants were asked to describe the mental images that they had generated for the word list and to rate the effectiveness of the mental imagery strategy.

In the second and third training sessions, participants saw a brief review of the examples of mental imagery that were presented in the first session and were then asked to practice the mental imagery strategy while completing memory tests with the same structure as the tests used in the first training session. The number of words (abstract and concrete) per study list and encoding time per word is shown in Table 1.

Table 1. Design of recall tests used in mental imagery training sessions.

Imagery Training Practice List	Training Session	Total Number of Study Words	Number of Abstract Study Words	Encoding Time Per Word
1	1	10	1	Controlled by Participant
2	1	10	2	3 seconds
3	2	10	2	3 seconds
4	2	10	3	3 seconds
5	2	10	3	2 seconds
6	2	10	3	2 seconds
7	2	12	3	2 seconds
8	2	12	4	2 seconds
9	3	12	4	3 seconds
10	3	12	4	2 seconds
11	3	14	4	2 seconds
12	3	14	5	2 seconds
13	3	16	5	2 seconds
14	3	16	6	2 seconds

5.5. Working Memory Training

5.5.1. Adaptive N-back Task

In the n-back task, single letters appeared sequentially on the screen and participants were required to indicate with a button press whether the current letter had appeared *n* items previously or not. For example, if subjects were shown the sequence A-B-C-B in a 2-back task, they would indicate that the second "B" was a target because it matched the letter that had appeared two letters back. They would respond „non-target' to the other items in the sequence. In order to train resolution of conflict/interference, the version of the task used in the current experiment (see also Novick et al. submitted) included lure trials. The lures were letters that appeared before (n-1) or after (n+1) the nth-back item (Kane, Conway, Miura, & Colflesh, 2007; see also Gray et al., 2003). For example, in the sequence *A-B-A-C-D* in a 3-back task, the second *A* is a lure (an *n*-1 lure) because it repeats a letter that appeared previously but not the correct number of letters back for the 3-back task. Subjects would have to resolve the interference of these familiar but non-target stimuli. Participants were presented with three lure conditions that corresponded with three levels of difficulty (no lures, n+1 lures only and both n+l and n-1 lures). The difficulty level of the task increased as per participants' performance. When participants achieved at least 85% accuracy at the current level (of *n* and lure difficulty level), difficulty was increased, first by increasing lure difficulty and then by increasing the *n*. If accuracy fell below 65%, task difficulty decreased, again first lure difficulty was decreased and then *n*. Task difficulty, therefore, represented both the value of *n* and the lure difficulty level.

5.5.2. Symmetry Span Task.

The symmetry span task required participants to remember the locations of a sequence of blocks that appeared in a 4x4 grid, in the order in which they were presented. The presentation of the blocks was interleaved with the presentation of a design on a different grid. Participants had to determine if the design was symmetrical across the vertical axis. Participants practiced the two task separately and then performed both in the dual-task phase. At the end of a series of these presentations, participants reported the remembered blocks by clicking on their positions, in order, on a blank grid.

The difficulty of this task was adjusted by changing the number of blocks that the participant must remember. The length of each series of memory items was determined by the participant's performance on the last set of judgments. Performance was re-evaluated after every four memory responses. If the participant got three or more correct, the sequence length increased by one. Conversely, if performance fell below two correct, the sequence length decreased by one. All participants started with a sequence length of three blocks.

5.6. Behavioral Results

5.6.1. Mental Imagery Training

Twenty-three of the 24 participants completed all three of the imagery training sessions and one participant completed only two of the training sessions. The average number of words recalled by the participants remained fairly consistent across the 14 memory tests used in the imagery training session, even as the encoding task became more difficult (longer word lists, shorter encoding times, more abstract words). Participants recalled an average of 9.04 words on the first memory test and an average of 8.14 words on the last memory test. As the memory tests became more difficult, participants reported that it became more difficult to create mental images for the word lists and they felt that the imagery strategy was less effective for the more difficult lists. The participants' ratings of how easy it was to create mental images (where 1 was easy and 5 was difficult) increased from an average of 2.96 for the first memory test to an average of 4.0 for the last, most difficult test. Similarly, the participants' ratings of the effectiveness of the memory strategy (where 1 is not effective and 5 is very effective) decreased from an average of 4.17 for the first memory test to an average of 2.5 for the last memory test.

5.6.2. Working Memory Training

Twenty-four of the 25 participants in the WM training group completed at least 12 of the 14 WM training sessions and one participant completed 9 of the training sessions. The participants who completed at least 12 of the WM training sessions were included in the analysis of the WM training. On average, the participants' performance improved across the training sessions for both training tasks. During the first training session, the participants had an average N-back level of 1.81 and an average symmetry span difficulty level of 3.77. On the 12th training session, the participants' average N-level was 4.23 and their average symmetry span difficulty level was 5.43. However, there was a great deal of variability across participants. The average N-level

achieved by each participant on the 12th training session ranged from 1 to 8.83. Similarly, the average level of difficulty achieved by each participant for the 12th session of the symmetry span task ranged from 3.18 to 7.33.

5.6.3. Baseline Tasks

5.6.3.1. Recognition Memory Task.

The participants' recognition memory performance (average proportion correct for each condition) is shown in Table 2. For analysis of the pre-training session, the performance was collapsed across training groups. The average performance of all participants on the pre-training recognition test is shown in Figure 24. Paired t-tests showed that the participants' memory performance was significantly better for words that were repeated or quizzed relative to words that were studied only once (all t's > 10.31, all p's < 0.001). In addition, there was a spacing effect for both the repeated and the quizzed study items such that participants had significantly better memory performance for long lag items than for short lag items ($t(73) = 2.72$, $p < 0.01$ for repeated words; $t(73) = 5.13$, $p < 0.001$ for quizzed words).

The crucial comparison for examining the effects of the memory training techniques was the difference between pre-training and post-training recognition memory performance for the three training groups. The differences for each group (post-training accuracy minus pre-training accuracy on each of the memory test conditions) are shown in Figure 25. Paired t-tests were used to assess each group's change in performance for each test condition. The t-tests showed that the control group's performance in the pre-training and post-training sessions did not differ significantly for any condition (all t's < 1.12, all p's > 0.27). The participants in the mental imagery training group performed significantly better in the post-training session for the once-presented words ($t(23) = 1.80$, $p = 0.04$), short-lag repeated words ($t(23) = 2.40$, $p = 0.01$) and for the short-lag quizzed words ($t(23) = 2.17$, $p = 0.02$). The participants in the working memory training group performed significantly worse in the post-training session relative to the pre-training session for the long-lag repeated items ($t(24) = 2.81$, $p < 0.01$), the short-lag quizzed words ($t(24) = 2.26$, $p = 0.02$) and the long-lag quizzed words ($t(24) = 4.55$, $p < 0.01$). Their performance was marginally worse the once-presented words ($t(24) = 1.65$, $p = 0.06$).

For each test condition, a one-way ANOVA was used to compare the change in performance (post-training accuracy minus pre-training accuracy) across the three groups. Although the three training groups' change in performance did not differ significantly for correct rejections of new, unstudied items ($F(2, 71) = 0.30$), there were significant differences in the participants' change in hit rates in every other condition (once-presented words, $F(2, 71) = 3.50$, $p = 0.04$; short-lag repeated words, $F(2, 71) = 3.33$, $p = 0.04$; long-lag repeated words, $F(2, 71) = 3.60$, $p = 0.03$; short-lag quizzed words, $F(2, 71) = 4.23$, $p = 0.02$; long-lag quizzed words, $F(2, 71) = 8.12$, $p < 0.01$).

Table 2. Behavioral performance on recognition memory task.

	Condition	Training Group	Pre-training Proportion Correct	Post-training Proportion Correct	Change in Performance (Post – Pre)
Words quizzed during study session	Short lag quizzes	Control Group	0.92	0.94	0.02
		Mental Imagery	0.96	0.98	0.02
		Working Memory	0.92	0.90	-0.01
	Long lag quizzes	Control Group	0.69	0.71	0.02
		Mental Imagery	0.74	0.80	0.06
		Working Memory	0.70	0.63	-0.07
	New words quizzed	Control Group	0.88	0.93	0.05
		Mental Imagery	0.91	0.93	0.02
		Working Memory	0.94	0.93	0.00
Words tested during subsequent memory test	Once-studied words	Control Group	0.84	0.83	-0.01
		Mental Imagery	0.83	0.84	0.01
		Working Memory	0.84	0.82	-0.02
	Short lag repetitions	Control Group	0.47	0.50	0.03
		Mental Imagery	0.57	0.64	0.07
		Working Memory	0.47	0.42	-0.05
	Long lag repetitions	Control Group	0.60	0.62	0.02
		Mental Imagery	0.68	0.76	0.08
		Working Memory	0.61	0.58	-0.03
	Short lag quizzes	Control Group	0.60	0.62	0.02
		Mental Imagery	0.75	0.78	0.03
		Working Memory	0.64	0.55	-0.09
	Long lag quizzes	Control Group	0.65	0.64	-0.01
		Mental Imagery	0.72	0.77	0.05
		Working Memory	0.67	0.61	-0.06
	New words	Control Group	0.68	0.66	-0.02
		Mental Imagery	0.78	0.81	0.03
		Working Memory	0.77	0.65	-0.12

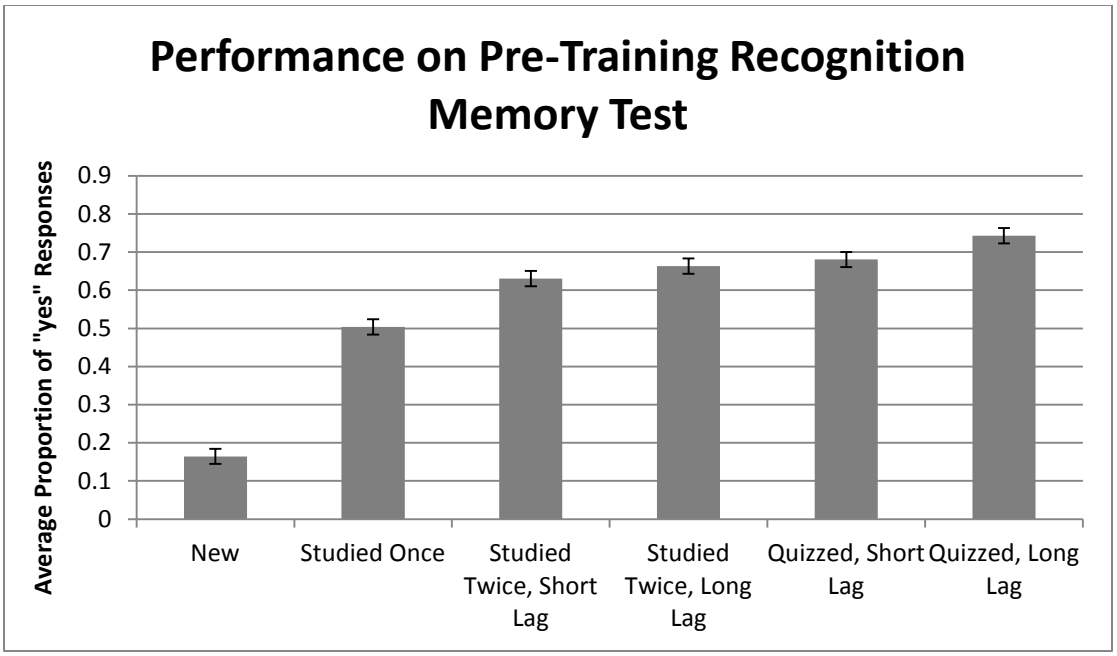


Figure 24. Average proportion of "yes" responses for each condition on the pre-training recognition memory test.

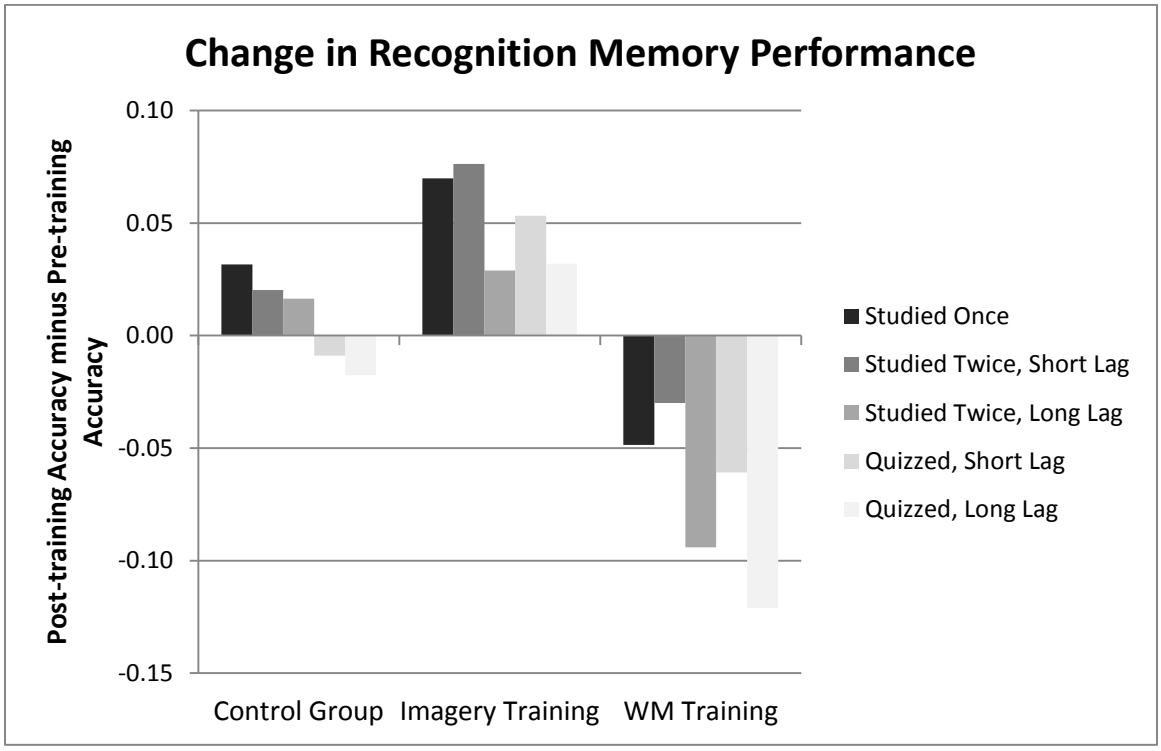


Figure 25. Difference between post-training memory performance (% correct) and pre-training memory performance for each condition on the recognition memory test, compared across training groups.

5.6.3.2. Memory Error Task.

The participants' performance on each test condition in the memory error task is shown in Table 3. Paired t-tests were used to assess each group's change in performance between the pre-training and post-training sessions and showed that none of the groups had a significant change in performance for any condition (all t's < 1.37, all p's > 0.09). For each test condition, a one-way ANOVA was used to compare the change in performance (post-training accuracy minus pre-training accuracy) across the three training groups. There were no significant differences between the groups (all F's < 1.45, all p's > 0.24).

Table 3. Behavioral performance on memory error task

	Training Group	Pre-training Performance	Post-training Performance	Change in Performance (Post – Pre)
Old items (Hits)	Control Group	0.88	0.87	-0.01
	Mental Imagery	0.91	0.89	-0.02
	Working Memory	0.86	0.82	-0.04
New items (Correct Rejections)	Control Group	0.61	0.59	-0.02
	Mental Imagery	0.74	0.77	0.02
	Working Memory	0.66	0.61	-0.05
Semantic Lures (False Alarms)	Control Group	0.27	0.25	-0.02
	Mental Imagery	0.25	0.26	0.01
	Working Memory	0.28	0.26	-0.02
Conjunction Lures (False Alarms)	Control Group	0.28	0.30	0.02
	Mental Imagery	0.36	0.32	-0.04
	Working Memory	0.30	0.28	-0.02

5.6.3.3. Associative Memory Task.

The participants' performance on each test condition in the associative memory task is shown in Table 4. Paired t-tests were used to assess each group's change in performance between the pre-training and post-training sessions. For the control group, there were no significant differences between pre-training performance and post-training performance (all t's < 1.15, all p's > 0.13). The participants in the mental imagery training group had significantly higher hit rates for the intact picture pairs after training ($t(23) = 2.81$, $p < 0.01$) as well as significantly lower false alarm rates for the rearranged picture pairs ($t(23) = 3.53$, $p < 0.01$) and significantly higher correct identifications of rearranged picture pairs ($t(23) = 4.02$, $p < 0.01$). The participants in the working memory training group also had significantly lower false alarm rates for the rearranged

picture pairs after training ($t(24) = 1.95$, $p = 0.03$) but their post-training performance did not significantly differ from their pre-training performance for any other test condition.

Table 4. Behavioral performance on associative memory task

	Response	Training Group	Pre-training Performance	Post-training Performance	Change in Performance (Post – Pre)
Intact Picture Pairs	Q1: Yes Q2: Yes (Correct)	Control Group	0.67	0.69	0.02
		Mental Imagery	0.68	0.78	0.09
		Working Memory	0.67	0.70	0.03
	Q1: Yes Q2: No (Miss)	Control Group	0.21	0.19	-0.02
		Mental Imagery	0.24	0.18	-0.07
		Working Memory	0.20	0.19	-0.01
	Q1: No Q2: No (Miss)	Control Group	0.12	0.12	0.00
		Mental Imagery	0.07	0.04	-0.03
		Working Memory	0.13	0.10	-0.02
Rearranged Picture Pairs	Q1: Yes Q2: Yes (False Alarm)	Control Group	0.24	0.26	0.02
		Mental Imagery	0.30	0.18	-0.13
		Working Memory	0.23	0.18	-0.05
	Q1: Yes Q2: No (Correct)	Control Group	0.65	0.61	-0.03
		Mental Imagery	0.63	0.78	0.15
		Working Memory	0.64	0.68	0.04
	Q1: No Q2: No (Miss)	Control Group	0.11	0.13	0.02
		Mental Imagery	0.07	0.05	-0.02
		Working Memory	0.13	0.14	0.01
New Picture Pairs	Q1: Yes Q2: Yes (False Alarm)	Control Group	0.02	0.01	0.01
		Mental Imagery	0.01	0.01	0.00
		Working Memory	0.01	0.01	0.00
	Q1: Yes Q2: No (False Alarm)	Control Group	0.05	0.03	-0.02
		Mental Imagery	0.06	0.03	-0.03
		Working Memory	0.03	0.04	0.01
	Q1: No Q2: No (Correct)	Control Group	0.92	0.95	0.03
		Mental Imagery	0.93	0.96	0.03
		Working Memory	0.96	0.95	-0.01

For each test condition, a one-way ANOVA was used to compare the change in performance (post-training accuracy minus pre-training accuracy) across the three groups. There was a

significant difference between the groups for the change in false alarm rates ($F(2, 71) = 4.42$, $p = 0.02$) and for the change in the number of correct identifications of rearranged picture pairs ($F(2, 71) = 6.44$, $p < 0.01$). However, the groups did not differ significantly in their changes in response patterns to any other test condition.

5.6.3.4. Deese-Roediger-McDermott Task.

Due to a technical problem, data from two of the participants from the control group could not be included in the analysis of the DRM task. The results for the remaining participants are shown in Table 5. Paired t-tests were used to assess each group's change in performance between the pre-training and post-training sessions. Participants in the imagery training group had significantly higher hit rates after training ($t(23) = 1.79$, $p = 0.04$), but there were no significant changes in performance for any of the other training groups or conditions (all t's < 1.39, p's > 0.09). In addition, there were no significant changes in the percentages of "Type A" and "Type B" memories for any of the groups (all t's < 1.48, p's > 0.08). For each test condition, a one-way ANOVA was used to compare the change in performance (post-training accuracy minus pre-training accuracy) across the three training groups. There were no significant differences between the groups (all F's < 1.03, all p's > 0.36).

Table 5. Behavioral performance on DRM task

	Training Group	Pre-training Performance	Post-training Performance	Change in Performance (Post – Pre)
Old items (Hits)	Control Group	0.48	0.52	0.04
	Mental Imagery	0.60	0.66	0.06
	Working Memory	0.46	0.49	0.03
Unrelated new items (Correct Rejections)	Control Group	0.90	0.90	0.00
	Mental Imagery	0.90	0.91	0.01
	Working Memory	0.94	0.92	-0.01
DRM Lures (False Alarms)	Control Group	0.58	0.56	-0.01
	Mental Imagery	0.66	0.69	0.02
	Working Memory	0.56	0.56	0.00

5.6.3.5. Rotation Span Task.

Two participants were excluded from the analysis of the rotation span task (one from the control group and one from the WM training group) due to failure to complete the post-training test. The results for the remaining participants are shown in Table 6. Paired t-tests were used to assess each group's change in performance between the pre-training and post-training sessions. All three training groups performed significantly better during the post-training session (all t's > 3.62, p's < 0.01). However, a one-way ANOVA comparing the change in performance across all

three training groups showed that there were no significant differences between the groups ($F(2, 69) = 0.49$, $p = 0.61$).

Table 6. Performance on Rotation Span Task

Training Group	Pre-training Performance	Post-training Performance	Change in Performance (Post – Pre)
Control Group	0.60	0.70	0.10
Mental Imagery	0.57	0.64	0.08
Working Memory	0.57	0.68	0.12

5.6.3.6. Listening Span Task.

Five participants (four from the control group and one from the WM training group) were excluded from the analysis of the listening span task due to a problem with the presentation of the sound files during the pre-training session. The results from the remaining participants are shown in Table 7. Paired t-tests were used to asses each group's change in performance between the pre-training and post-training sessions. All three groups performed significantly better during the post-training session (all t's > 1.87, p's < 0.04). However, a one-way ANOVA showed that the three groups' change in performance were not significantly different from one another ($F(2, 66) = 1.69$, $p = 0.19$).

Table 7. Performance on Rotation Span Task

Training Group	Pre-training Performance	Post-training Performance	Change in Performance (Post – Pre)
Control Group	34.67	42.76	8.10
Mental Imagery	37.75	43.88	6.13
Working Memory	41.04	44.04	3.00

5.7. Electrophysiological Results

Since the most notable difference between the training groups were in the recognition memory task, we focused the ERP analysis on that task.

Within the grand average waveforms, the N400 was measured by computing the mean amplitude of the ERPs in a time window from 300-500 ms post stimulus onset. The LPC was measured by computing the mean amplitude of the ERPs in a time window from 500-800 ms post stimulus onset. Repeated measures ANOVAs were used to test the results using seven electrodes centered

over the posterior midline of the scalp (Z7, Z8, Z9, L7, L8, R7, and R8). All effects are significant at or above the $p = 0.05$ level unless otherwise specified. For clarity, all plots show electode Z7, which was located on the vertex of the scalp (equivalent to electrode position Cz in 10-20 systems).

5.7.1. Recognition Memory Task.

5.7.1.1. Pre-training ERPs.

For the ERPs recorded during study, the first presentations of all studied words were compared to the repeated and quizzed words. Figure A shows the ERPs elicited at study by the repeated words and Figure B shows the ERPs elicited by the quizzed words. When the repeated words were compared to the first presentations of the words, there was a significant reduction in the amplitude of the N400 component for words that were repeated at short lags ($F(1,72) = 7.35$, $p < 0.01$) but not for words that were repeated at long lags ($F(1,72) = 2.09$, $p < 0.15$). These data indicate that there was an N400 repetition effect for repeated words at short lags, but that the effect was attenuated at long lags.

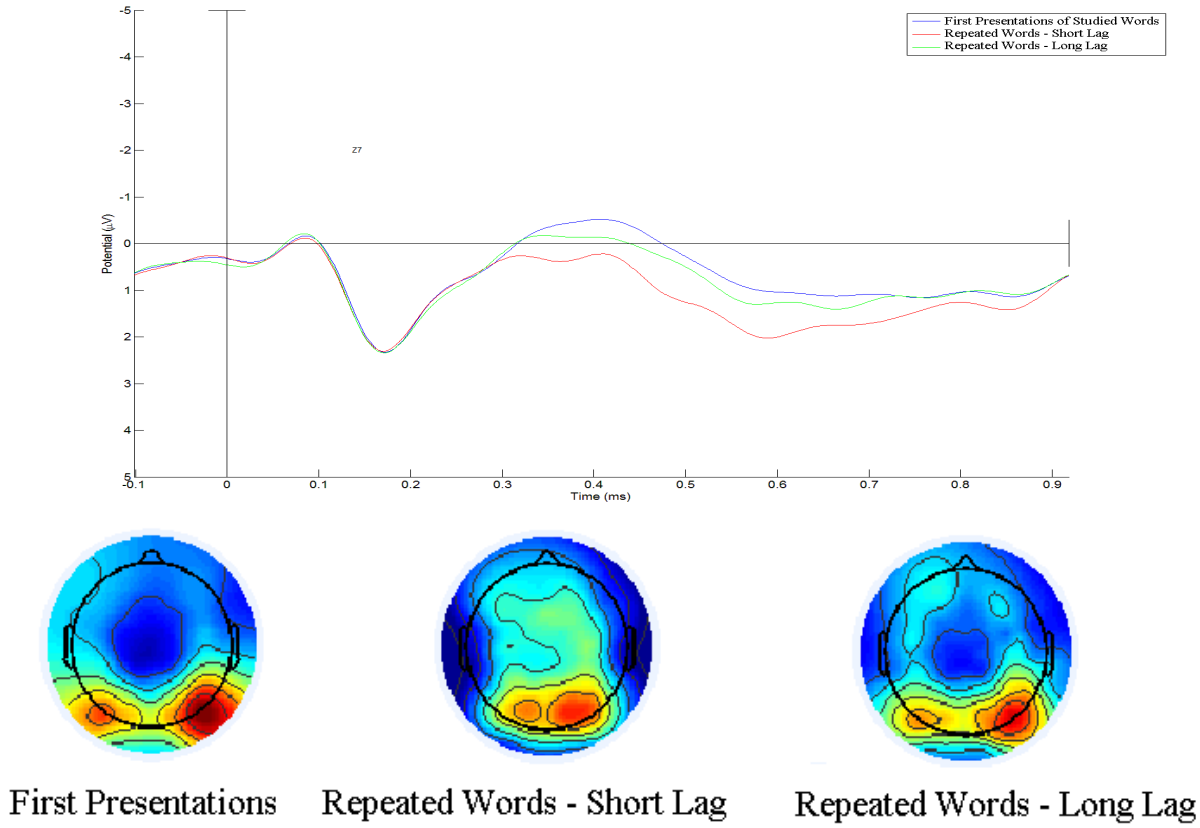


Figure 26. N400 time window scalp maps and ERPs recorded at electrode Z7 for the first presentation of studied words and short- and long-lag repetitions.

When the words that were quizzed during the study session were compared to the first presentations of studied words, there was a significant reduction in the amplitude of the N400 component for the words that were quizzed at both short lags ($F(1,72) = 6.18$, $p < 0.02$) and long lags ($F(1,72) = 5.32$, $p < 0.03$). In addition, there was a larger LPC component for the quizzed words at both short ($F(1,72) = 17.68$, $p < 0.01$) and long lags ($F(1,72) = 11.98$, $p < 0.01$).

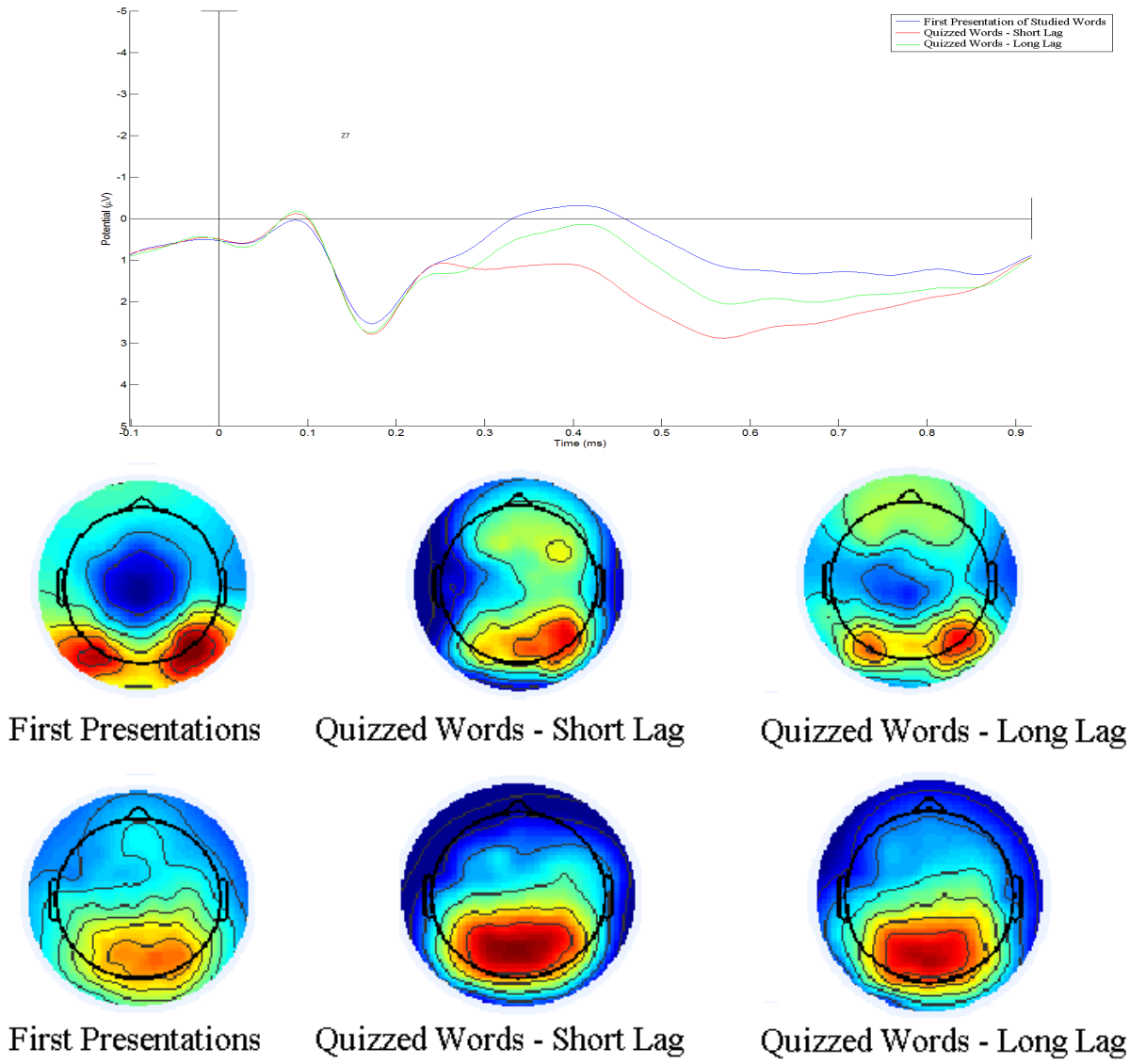


Figure 27. ERPs recorded at electrode Z7 for the first presentation of studied words and short- and long-lag quizzes, N400 time window scalp maps (center) and LPC time window scalp maps (bottom).

These results indicate that, as expected, participants' processing of the second presentation of a studied word was facilitated when there was a short lag between the first and second presentations. There was a larger LPC when the words were quizzed, indicating that participants were searching their memory for the quizzed words.

The traditional Dm effect was calculated by backsorting the ERPs elicited by the critical words based on subsequent memory for those words. In order to study the impacts of the repetitions and quizzes on subsequent memory, Dm-R effects were also calculated for the second presentation of repeated and quizzed words. For all conditions the Dm and Dm-R effects were largest over posterior channels. Analyses of the mean amplitude of the waveforms were carried out in the N400 time window from 300-500 ms and in the LPC time window from 500-800 ms.

For the words that were presented only once, there was a significant Dm effect, with subsequently remembered words having a more positive waveform than subsequently forgotten words, in the LPC time window (shown in Figure 28). There were no significant Dm or Dm-R effects for any other condition.

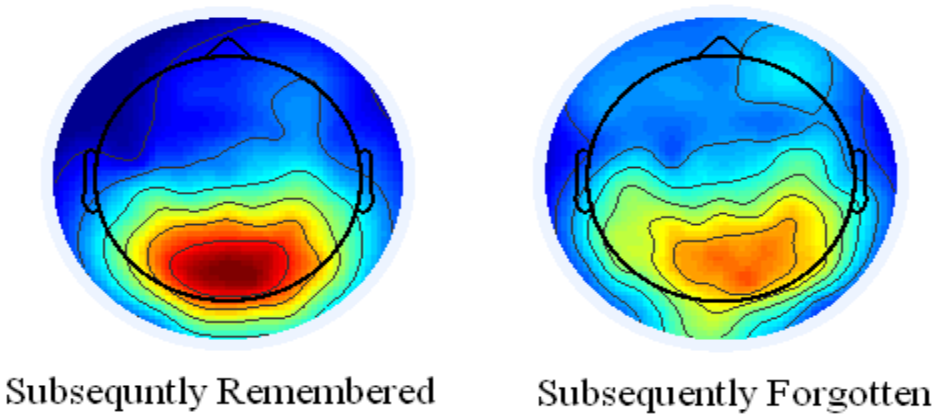


Figure 28. LPC time window scalp maps showing the Dm Effect for once-presented words.

5.7.1.2. Post-training ERPs.

The ERPs recorded in the post-training session showed the same pattern of results as the pre-training session for all three memory training groups. In order to compare across groups, difference waves were calculated for the repetition effects at study (first presentations minus second presentations).

The comparisons of the difference waves across groups are shown in Figure 29 for the short lag repeated words, in Figure 30 for the long lag repeated words, in Figure 31 for the short lag quizzes and in Figure 32 for the long lag quizzes. For the long lag repeated words, the difference between the first presentations of the words and the repeats was not significant either before or after training, producing near-zero difference waves. These are reflected in the scalp maps in Figure 30. There were no meaningful differences between the pre- and post-training sessions for any of the training groups in that condition. However, there were some differences related to training in the other conditions.

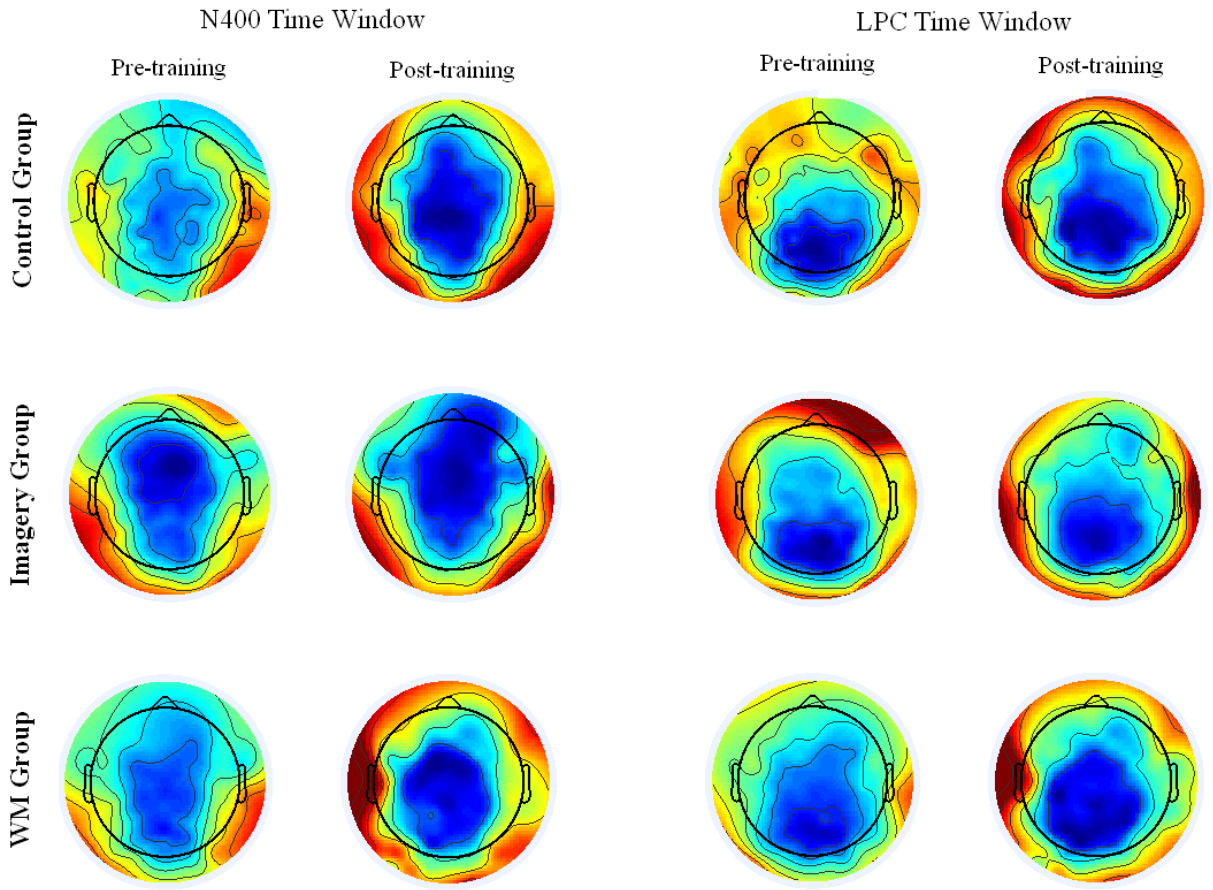


Figure 29. Difference waves for short lag repeated words.

For the participants in the mental imagery training group, the negativity of the difference waves in the N400 time window had a more anterior distribution after training for the short lag repeated words and short lag quizzes. In addition, their difference waves were smaller after training in the LPC time window for the short lag quizzes.

For the control and WM training groups, the difference between the first presentations of training words and the short lag repetitions was larger after training in the N400 time window, as shown by the increased negativity in the post-training scalp maps. The WM training group had smaller difference waves in both the N400 and LPC time windows for the short lag quizzes.

The difference wave results indicate that the participants in the imagery training group are processing the short lag items differently after training. The more anterior distribution of the difference waves could reflect increased use of visualization during encoding. The smaller difference waves for the WM training group in the quizzed conditions may indicate that the participants in that group are expending less effort in retrieving information from memory for the quizzes. This explanation could account for their poorer behavioral performance at test.

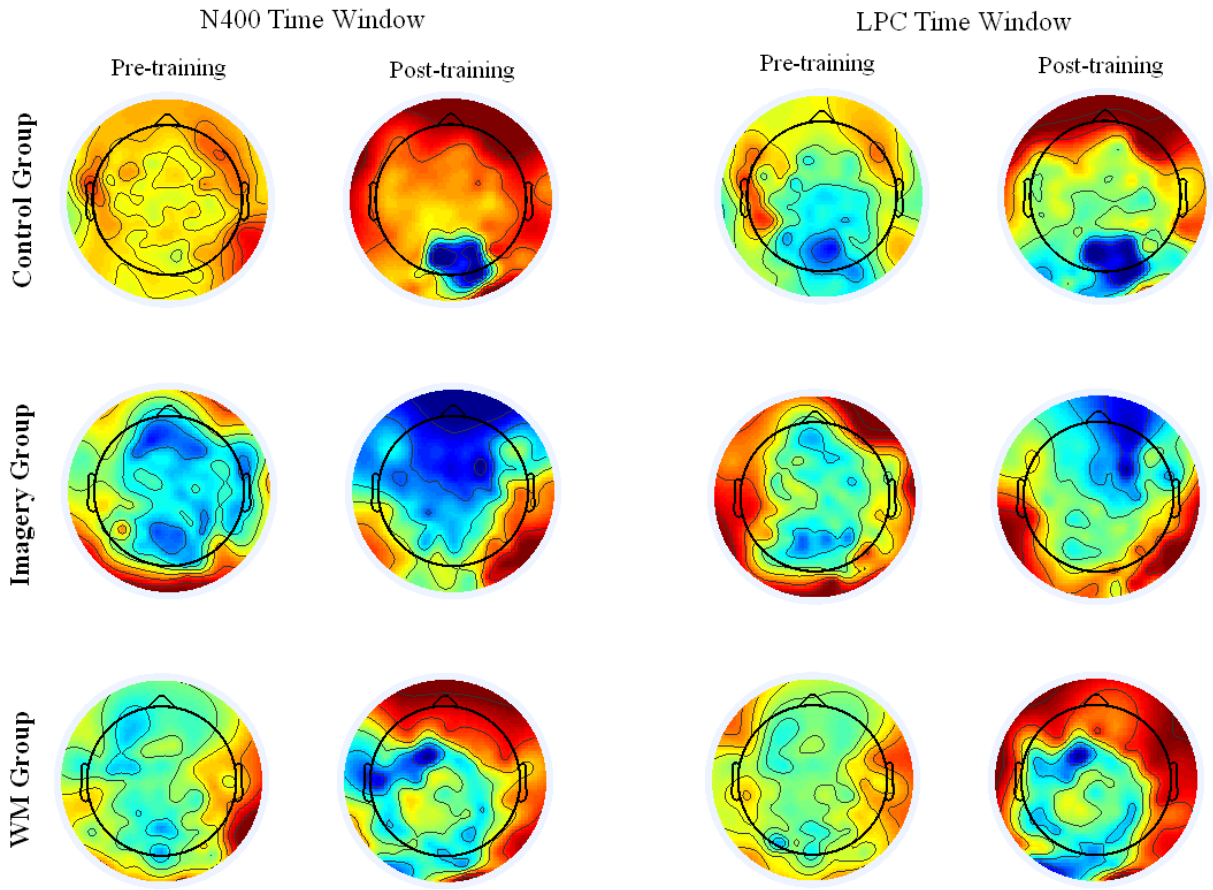


Figure 30. Difference waves for long lag repeated words.

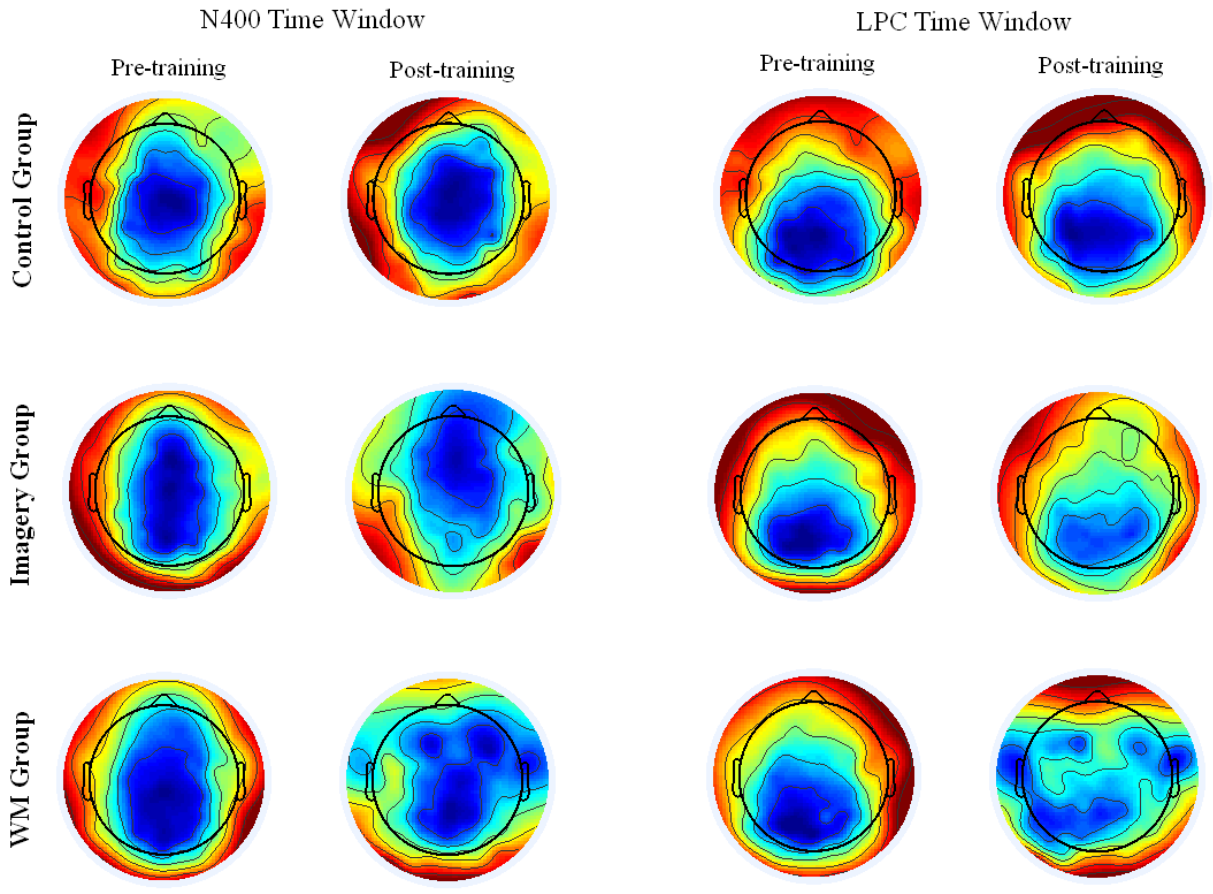


Figure 31. Difference waves for short lag quizzed words.

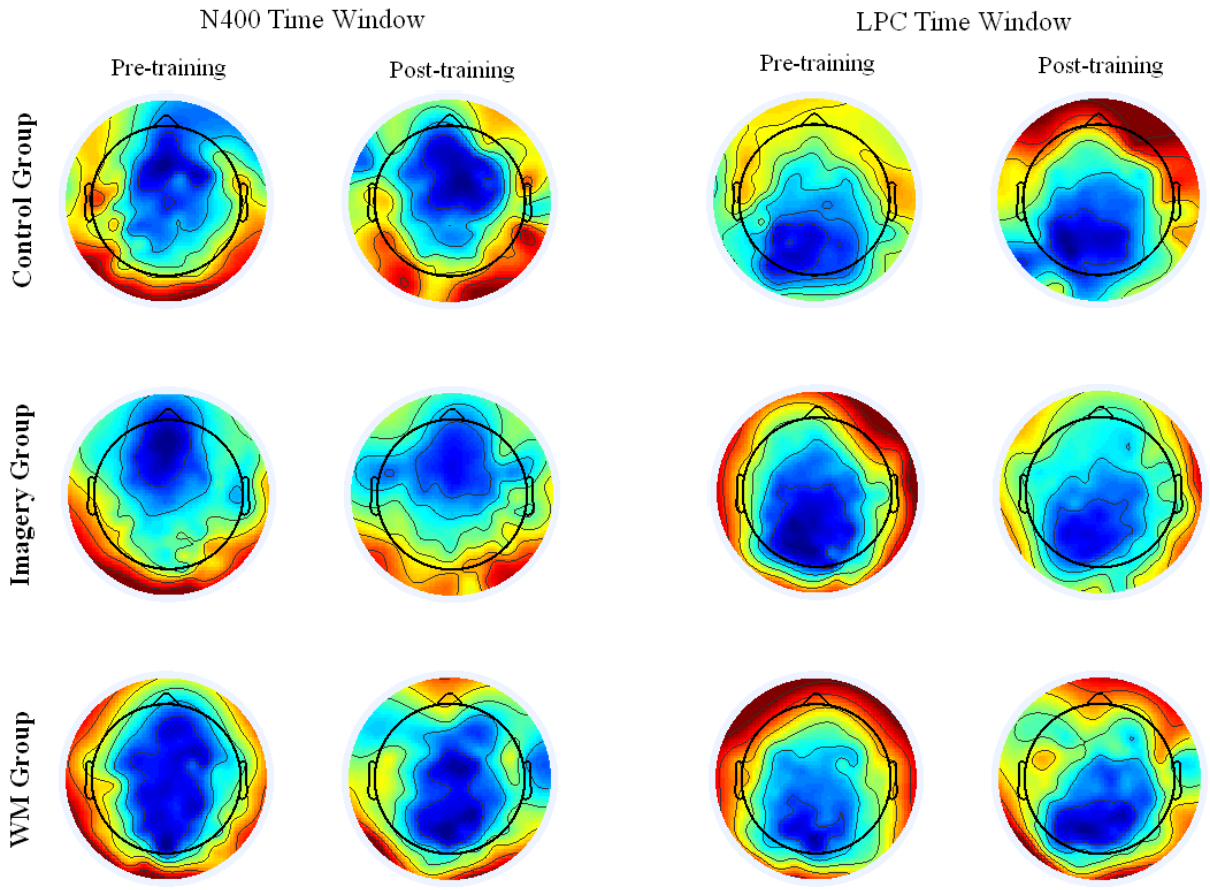


Figure 32. Difference waves for long lag quizzed words.

5.8. Discussion

The results of the memory training study indicate that using an appropriate memory strategy is crucial for performing well on a variety of memory tasks. While the participants who practiced using a memory strategy improved their performance on a variety of memory tasks, the participants who completed the working memory training did not improve their performance in most cases. In fact, their performance decreased after training on several conditions in the recognition memory task, a task in which a memory strategy would be helpful.

The ERP results indicate that the participants in the working memory training group may have put less effort into retrieving information from memory when they were being quizzed during study. Experiments 1 and 2 demonstrated the effectiveness of using a retrieval practice strategy and the participants in the working memory training group appear to be using a less effective strategy after training, as indicated by their smaller LPCs in the quiz conditions. The nature of the working memory training may lead participants to give up on using a memory strategy, which then impairs their performance on memory tasks where a strategy would be beneficial.

In future work, we would like to investigate the differences between groups of participants who responded to the memory training compared to those who did not. The present study indicates that there are substantial individual differences in how people approach the memory tasks before training, how well they perform on the training sessions, and how the training changes their inherent use of memory strategies. Exploring this issue further could allow us to predict which people would benefit from memory training and what type of training would be most helpful for a specific individual.

6. REFERENCES

Allan, K., Rugg, & M. D. (1997). An event-related potential study of explicit memory on tests of cued recall and recognition. *Neuropsychologia, 35*, 387-397.

Anderson, J. E., & Holcomb, P. J. (2005). An electrophysiological investigation of the effects of coreference on word repetition and synonymy. *Brain and Language, 94*, 200-216.

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., & Treiman, R. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. http://elexicon.wustl.edu/, Washington University.

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In Ross, B.H., Benjamin A.S. (eds.) *The Psychology of Learning and Motivation: Skill and Strategy in Memory Use. Vol 48*, pp 175-223. Academic Press, London.

Benjamin, A.S., Bjork, R.A., & Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55-68.

Bentin, S., & Moscovitch, M. (1990). Psychophysiological indices of implicit memory performance. *Bulletin of the Psychonomic Society, 28(4),* 346-352.

Bentin, S., & Peled, B. S. (1990). The contribution of task-related factors to ERP repetition effects at short and long lags. *Memory and Cognition, 18*, 359-366.

Besson, M., & Kutas, M. (1993). The many facets of repetition: Cued-recall and event-related potential analysis of repeating words in same versus different sentence contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1115-1133.

Besson, M., Kutas, M., & Van Petten, C. (1992). An event-related potential (ERP) analysis of semantic congruity and repetition effects in sentences. *Journal of Cognitive Neuroscience, 4*, 132-149.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimatnura (Eds.), *Metacopition: Knowing about knowing*, pp. 185-205. Cambridge, MA: MIT Press.

Bjork, R.A. (1999). Assessing our own competence: Heuristics and Illusions. In Gopher D., Koriat A. (eds.) *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*. pp. 435-459. MIT Press, Cambridge, MA

Blajenkova, O., Kozhevnikov, M., & Motes, M. A. (2006). Object-Spatial imagery: A new self-report imagery questionnaire. *Applied Cognitive Psychology, 20*, 239-263.

Daneman, M., & Blennerhassett, A. (1984). How to assess the listening comprehension skills of prereaders. *Journal of Educational Psychology, 76*, 1372 -1381

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

Delorme, A. & Makeig, S. (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods, 134*, 9-21

Federmeier, K. D., & Benjamin, A. S. (2005). Hemispheris asymmetries in the time course of recognition memory. *Psychonomic Bulletin & Review, 12 (6),* 993-998.

Koriat, A., & Bjork, R.A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187-194.

Koriat, A., Bjork, R.A., Sheffer, L., & Bar, S.K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643-656.

Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Kutas, M., & Federmeier, K. D. (2000). Electophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences, 4, 463-470.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In Gruneberg, M. M., Morris, P. E., Sykes R. N. (eds.) *Practical Aspects of Memory*. pp. 625-632. Academic Press, London.

Laszlo, S., & Federmeier, K. D. (2007). Better the DVL you know: Acronyms reveal the contribution of familiarity to single word reading. *Psychological Science, 18*, 122-126.

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology, 64*, 17-24.

Matzen, L. E., & Benjamin, A. S. (2009). Remembering words not presented in sentences: How study context changes patterns of false memories. *Memory & Cognition, 37,* 52-64.

Nagy, M. E., & Rugg, M. D. (1987). Modulation of ERPs by word repetition: The eVects of inter-item lag. *Psychophysiology, 26*, 431–436.

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

Olichney, J. M., Van Patten, C., Paller, K. A., Salmon, D. P., Iragui, V. J., & Kutas, M. (2000). Word repetition in amnesia: Electrophysiological measures of impaired and spared memory. *Brain, 123*, 1948-1963.

Paller, K. A., & Gross, M. (1998). Brain potentials associated with perceptual priming versus explicit remembering during the repetition of visual word-form. Neuropsychologia, 36, 559-571.

Paller, K. A., & Kutas, M. (1992). Brain potentials during memory retrieval provide neuropsychological support for the distinction between conscious recollection and priming. *Journal of Cognitive Neuroscience, 4*, 375-391.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29*, 633-654.

Pujol, J., Deus, J., Losilla, J. M., Capdevila, A. (1999). Cerebral lateralization of language in normal left-handed people studied by functional MRI. *Neurology, 52*, 1038.

Roediger, H. L. & 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., Watson, J. M., McDermott, K. B. & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8(3),* 385-407.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology, 22*, 642–647.

Rugg, M. D. (1987). Dissociation of semantic priming, word and nonword repetition effects by event-related potentials. *Quarterly Journal of Experimental Psychology, 39A*, 123–148.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition, 18*, 367–379

Rugg, M. D., & Doyle, M. C. (1994). Event-related potentials and stimulus repetition in direct and indirect tests of memory. In Heinze, H. J., Munte, T. F., Mangun, G. R. (eds.) *Cognitive Electrophysiology.* pp. 124-138. Birkhauser, Boston, MA.

Schwartz, B.L., Benjamin, A.S., & Bjork, R.A. (1997). The inferential and experiential bases of metamemory. Current Directions in Psychological Science, 6, 132-137.

Shah P, & Miyake A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125*, 4–27.

Smith, M. E., & Halgren, E. (1989). Event-related potentials during lexical decision: Effects of repetition, word frequency, pronounceability pronounceability, and concreteness. *Electroencephalography and Clinical Neurophysiology, Supplement, 40*, 417–421.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., McIssac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience, 3*, 131-150.

Van Petten, C., & Senkfor, A. J. (1996). Memory for words and novel visual patterns: Repetition, recognition, and encoding effects in the event-related brain potential. *Psychophysiology, 33*, 491-508.

Wilson, M.D. (1988) The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, *20*, 6-11.

MATLAB version 7.11.0.584(R2010b). Natick, Massachusetts: The MathWorks Inc., 2010.

APPENDIX A: DEVELOPMENT OF COMPUTATIONAL MODELING METHODS FOR ERPS

Successful classification of single trial ERPs is desirable to assess the effectiveness of the intervention techniques in terms of neural processing. The majority of published ERP studies report results using "grand average" ERPs due to the low signal-to-noise ratio (SNR) of single trial ERP signals. The noise level for single trial ERPs is on order of 10's to 100's of microvolts, while grand average ERPs signals are on the order of 1-10 microvolts. Modeling methods that are robust to the noise of single trial ERP signals yet sensitive to the underlying signal levels of grand average ERPs are necessary to assess the effectiveness of the intervention techniques in an operationally relevant manner.

Of the published ERP studies that report results using single trial ERPs, many focus on classifying ERPs from conditions characterized by "large" (relatively high SNR) ERP signals like motor control functions (Ramoser, et. al, 1998) or the recognition of novel images in a series of repeated image presentations (Sellers, et. al., 2006). Modeling methods used to classify single trial ERPs in these publications include methods from signal processing (matched filters, Woody, 1967), and machine learning (i.e. regression models, Blankertz, et. al., 2002), however no single modeling method has proven to be effective for classification in the lower SNR domain characterizing learning and memory performance. Therefore, a screening approach has been applied to a variety of modeling methods chosen to span a broad range of model complexity.

The modeling methods chosen for evaluation were used to create feature vectors that were subsequently used in either of two classification algorithms, linear discriminant classification or decision tree classification. The chosen modeling methods were: Euclidean distance, statistical distance, Mahalanobis distance, principal components analysis (PCA), independent components analysis (ICA), frequency analysis, matched filters and a rational model based on known ERP characteristics associated with the cognitive processes under study. These methods can be broadly grouped into four categories based on the characteristics and assumptions of each method (see Table A1). The three distance metrics, Euclidean, statistical and Mahalanobis describe geometric relationships among the data. The two component analysis metrics describe the variance relationships among the data. Due to the underlying assumptions of the ICA method, it is not readily applicable to data containing inter-subject variations. Because of this limitation, ICA was not included in the first level of screening, but will be included in the second level (data containing only intra-subject variations). Frequency analysis and matched filter methods come from signal processing theory and treat the data as time series signals. The rational model approach is unique in its attempt to leverage expert domain knowledge from ERP analysis, however it should be noted that much of our understanding of ERP characteristics is based on grand average data, which may not readily transfer itself to subject average or single trail ERPs.

The form in which the ERP data were collected allows different data ensembling schemes to be applied to create multiple data sets with different SNR characteristics. The ensebling method which produces the highest SNR data set averages the single trial ERPs from each subject to form a set of 24 subject-average ERP signals (this set contains inter-subject variability, but with maximal signal averaging). The next highest SNR data set is formed by grouping single trial

ERP signals from each subject (this set contains intra-subject variability but with no signal averaging). The lowest SNR data set is formed by combining all single trial ERP signals from all subjects (this set contains intra and inter subject variability with no signal averaging). Modeling methods that performed best at each SNR level, as measured by percentage of samples correctly classified, will be carried forward to test at the next SNR level. For all cases, results will be obtained using subject-out or sample-out cross validation so as to more fairly characterize the modeling method's potential for use in the operational mode of true, forward classification on unknown samples.

Table A1. Categories of modeling methods

	Category			
	Geometric	**Variance**	**Signal Processing**	**Rational Model**
Method	Euclidean distance	Principal component analysis	Frequency analysis	ERP expertise
	Statistical distance	Independent component analysis	Matched filter	
	Mahalanobis distance			

Finally, the modeling methods will be tested on two different conditions from the ERP study. The first condition compares ERPs from first presentation of words that were studied and later quizzed (FPSTW) and second presentation of words that were quizzed at a long lag (SPSTWSLR). The second condition compares ERPs from words that were studied once and that were subsequently remembered or forgotten. Differences between the ERP grand averages in the first condition are more pronounced than in the second condition. These two conditions were chosen to test the effectiveness of the modeling methods as the difference between the two ERP signals became smaller. Figure A1 shows the relative SNR of each data treatment in arbitrary units. For this estimate of SNR, the signal is defined as the root-mean-square (RMS) value of the point to point differences between the grand average ERP signal for FPSTW (S1R) and the grand average ERP signal for SPSTWSLR (S1F). The noise is defined as the RMS of the variance of the set of subject average ERP signals (or set of single trial ERP signals). The relative SNR estimates were normalized to the subject average FPSTW data ensemble.

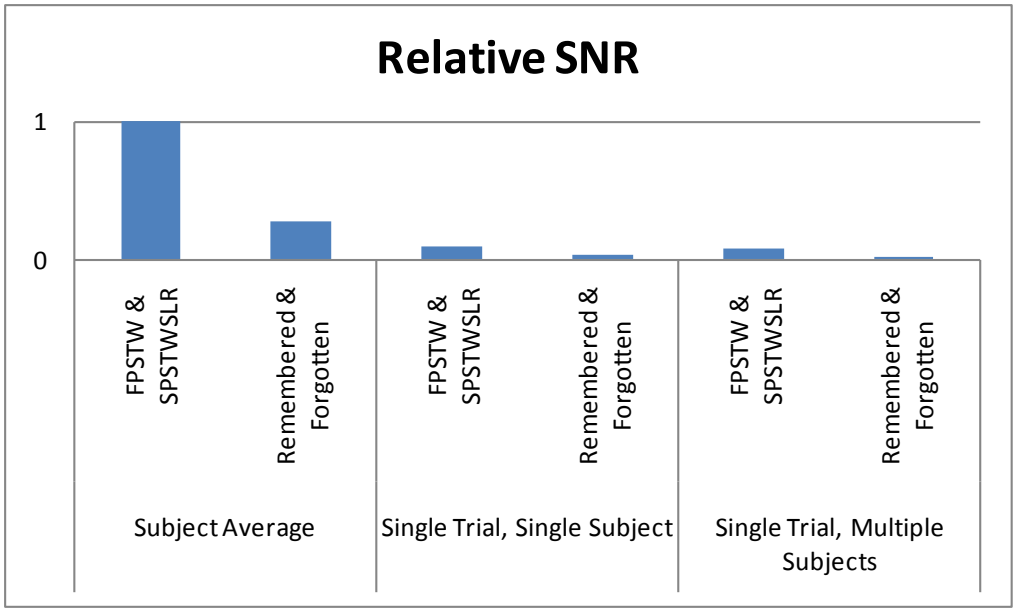


Figure A1. SNR of each data ensemble normalized to the highest SNR ensemble.

A1.1. Data Overview

Figures A2 and A3 show subject average and grand average ERPs for both tested conditions. The ERPs have been baseline corrected using the first 100 milliseconds (ms) prior to the stimulus presentation and low-pass filtered to remove frequencies higher than 25 Hz. For all modeling and classification work, the ERP signals were analyzed from the stimulus onset time (0 ms) to 800 ms after stimulus onset. Figure A4 shows the effect of the number of subjects included in the formation of the grand average ERP signals. It is clear from the figures that the individual subject variability is large compared to the difference between the two grand average ERP signals, thereby presenting a challenging classification problem.

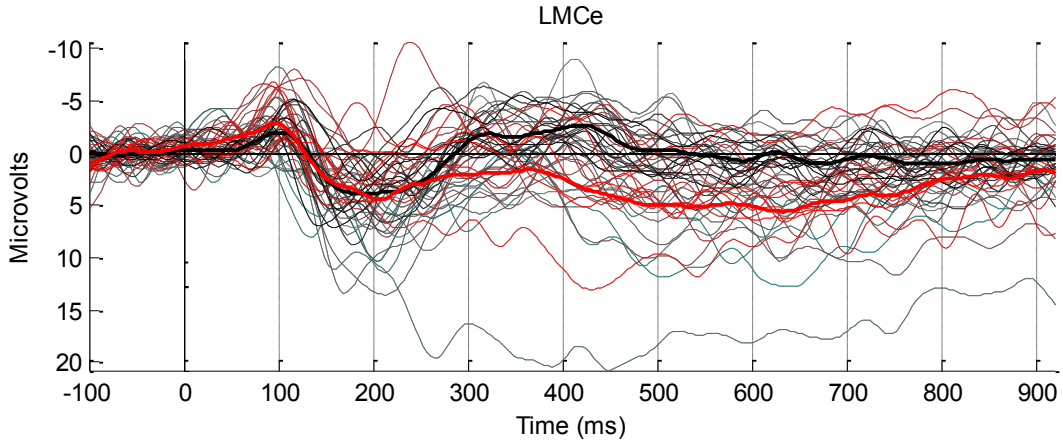


Figure A2. Subject average (dashed lines) and grand average (bold lines) ERP signals for FPSTW (red) & SPSTWSLR (black).

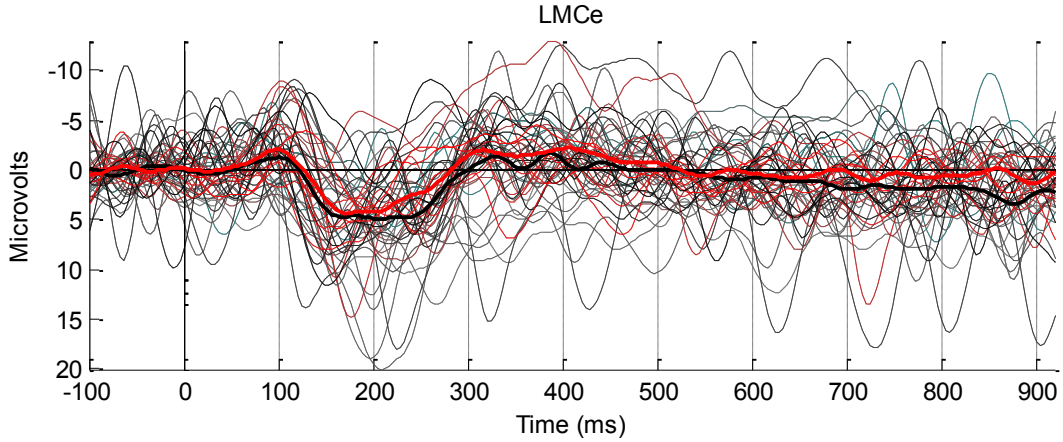
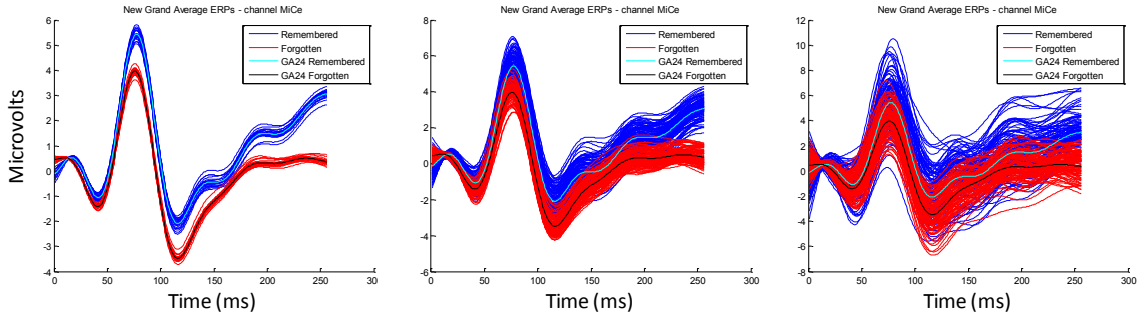


Figure A3. Subject average (dashed lines) and grand average (bold lines) ERP signals for S1R (red) & S1F (black).



a b c

Figure A4. Effect of number of subject used to form grand average ERP. The grand average from all 24 study subjects is shown in cyan and black. Grand averages formed by random selection of fewer than 24 subjects are shown in blue and red. (a. 24 random selections of 23 subjects to form grand average, b. 100 random selections of 15 subjects to form grand average, c. 100 random selections of 5 subjects to form grand average)

A1.2. Modeling Methods

The amplitude of the ERP signal at a point in time can be treated as a coordinate in a multidimensional space, thus each ERP signal (0 to 800 ms sampled at 250 Hz) represents a single point in a 201 dimensional space.

A1.2.1. Euclidean Distance

The Euclidean distance is an ordinary geometric distance between two points in this 201-dimensional space. This distance can be calculated from either the center of the pooled training data, or from the center of each training cluster (remembered or forgotten).

A1.2.2. Statistical Distance

The statistical distance is similar to the Euclidean distance except the ERP signals are standardized (z-score) prior to the distance calculation. The standardization parameters are defined only from the training data set, then, are applied to the unknown ERPs in subject-out cross validation.

A1.2.3. Mahalanobis Distance

The Mahalanobis distance is similar to the Euclidean distance in that it measures geometric distances from a group mean, however the Mahanlanobis distance also takes into account the correlation (or covariance) structures inherent in the data. For this work the Mahalanobis distance calculation was performed using the first five PCA factors of the training data as the basis set.

A1.2.4. Principal Component Analysis

Principle component analysis (PCA) transforms the ERP signals to a new, orthogonal basis set (principal components) based on maximizing the variance of the ERP signals associated with each new principle component. For this work, the scores of the first five PCA factors of the training data where used for classification.

A1.2.5. Independent Component Analysis

Independent component analysis (ICA) is similar to PCA, however instead of maximizing variance for each factor, ICA the independence of each factor.

A1.2.6. Frequency Analysis

The frequency analysis approach used for this work employed wavelet techniques to estimate the power present in the 10 Hz frequency band (relative to the power spectrum of the pre-stimulus ERP from 100 ms before the stimulus onset) in an eight millisecond (8 ms) moving window. The resulting vector of relative frequency powers over time where then processed using PCA, as described above, before the classification step.

A1.2.7. Matched Filter

In order to maximize the signal available for classification, a new approach of using matched filters was developed. The truth signals for each ERP state (i.e. remembered word or forgotten word) was defined as the grand average ERP signal constructed from the training data. Each unknown, subject-average ERP signal during subject-out cross validation was filtered by both truth signals. The new filtered signals where the appended and processed using PCA, as described above, before the classification step.

A1.2.8. Rational Model

The rational model method characterizes the ERP signals in three time windows identified to contain the most distinguishing features of the grand average ERP signals (see Figure A5). The peak height of the baseline corrected ERP signal over each of three time windows is used as the feature vector for classification. The time windows were defined as 150 to 250 ms after stimulus onset 350 to 450 ms after stimulus onset and 500 to 800 ms after stimulus onset.

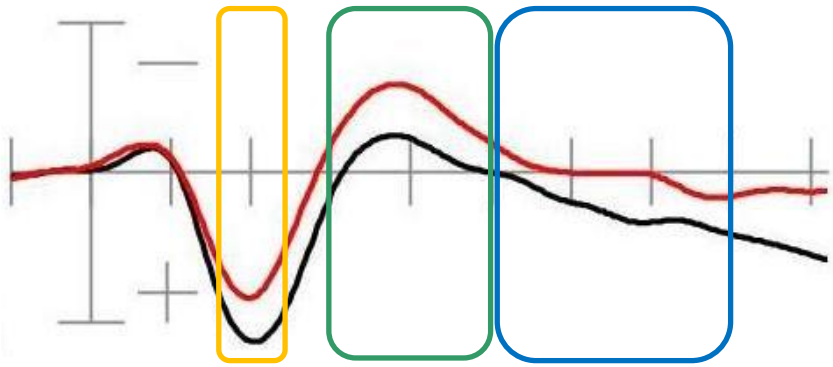


Figure A5. Illustration of time windows chosen for the rational model method.

A1.3. EEG Channel and Factor Selection

A1.3.1. Results

Results from the highest SNR level are shown in figures A6 and A7. Figure A6 shows the percentage of individual subject averages correctly classified in subject-out cross validation for the FPSTW and SPSTWSLR condition. Figure A7 shows the percentage of individual subject averages correctly classified using subject-out cross validation for the once-studied remembered and forgotten words. Since there are only two possible states (i.e. remembered or forgotten) for each sample, relative performance should be compared to a baseline of 50% correct classification (for a two state system, a random guess, or coin flip, for each sample would result in 50% correct classification).

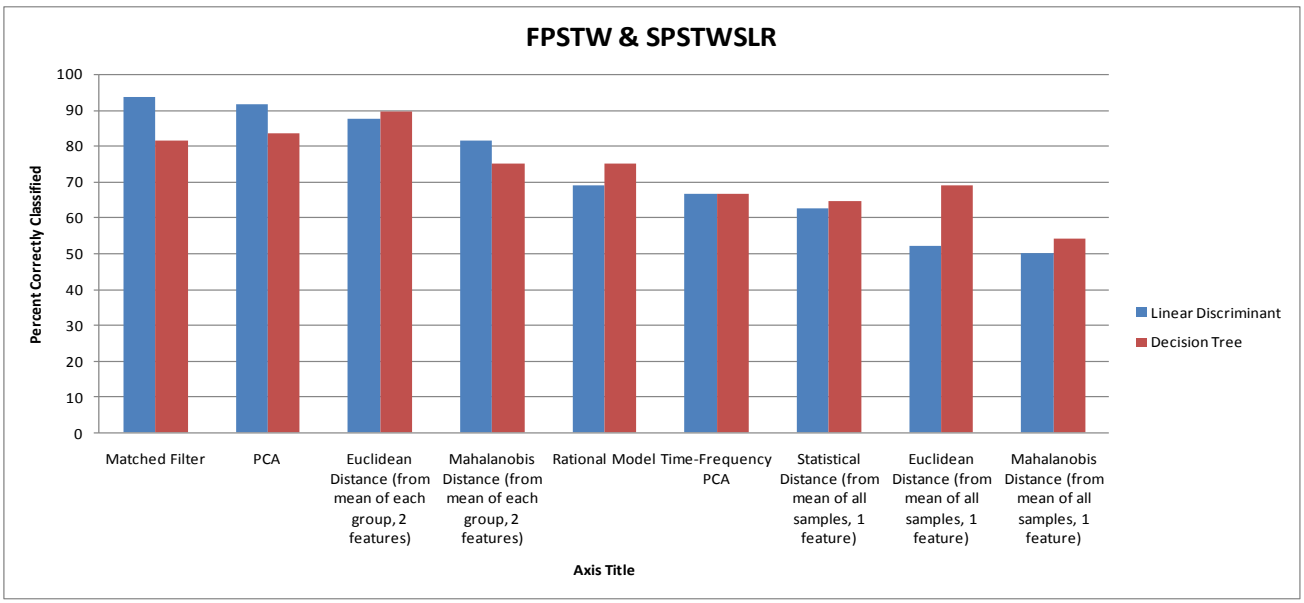


Figure A6. Percentage of individual subject averages correctly classified in subject-out cross validation for each modeling and classification method.

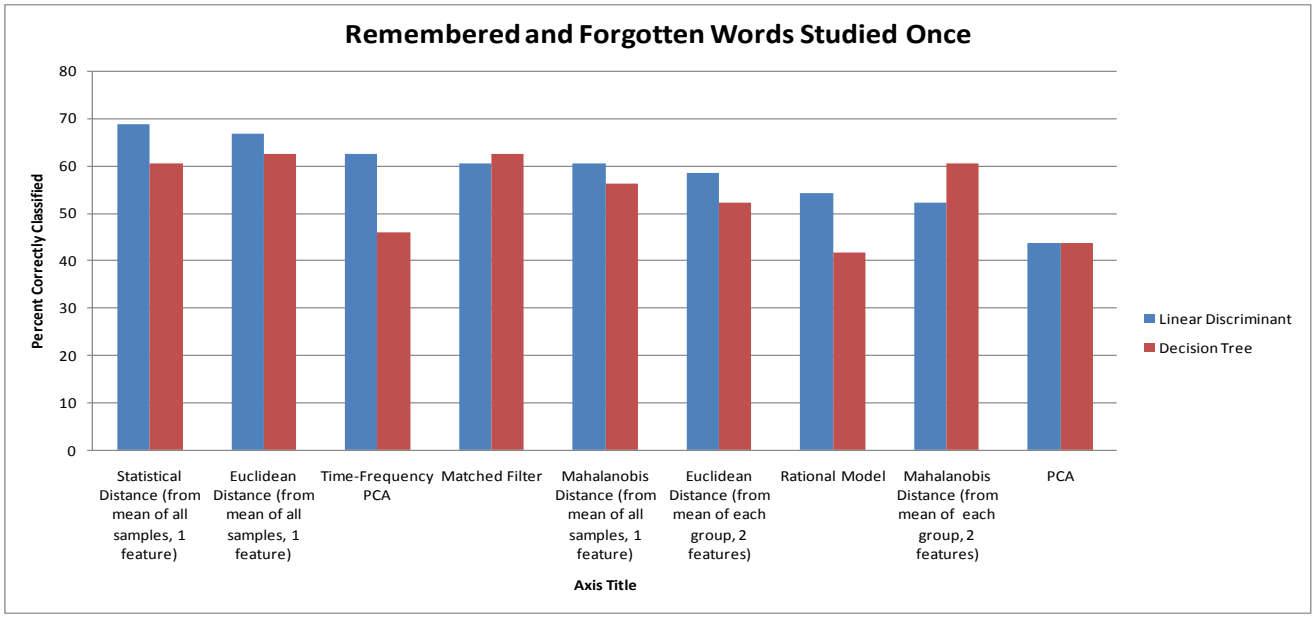


Figure A7. Percentage of individual subject averages correctly classified in subject-out cross validation for each modeling and classification method.

For the highest signal to noise case (Figure A6), the best performance comes from the matched filter modeling method and linear discriminant classification (~90% of the samples are correctly classified). For the next lower SNR case (S1F & S1R), the best performance comes from the statistical distance metric using linear discriminant classification (~70% of the samples are correctly classified). For this case the matched filter method also performed favorably compared to all other methods. It is somewhat surprising that one of the mulitvariate metrics (PCA,

Mahalanobis, Time-Frequency) did not produce results better than the simple statistical distance metric for the more difficult case of S1F & S1R. This may be due to the relatively small number of samples available, resulting in an insufficient representation of the underlying signal and the inter-subject variance present in this ERP study.

A1.4. References

Blankertz, B., G. Curio, and K. Müller, *Classifying single trial EEG:Towards brain computer interfacing*, in *Advances in Neural Inf. Proc.Systems (NIPS 01)*, T. iettrich, S. Becker, and Z. Ghahramani, Editors.2002. p. 157-164.

Ramoser, H., J. Müller-Gerking, and G. Pfurtscheller, *Optimal spatial filtering of single trial EEG during imagined hand movement.* IEEE Trans. Rehab. Eng, 1998(8): p. 441-446.

Sellers, E., et al., *A p300 event-related potential brain–computer interface (BCI): the effects of matrix size and inter stimulus interval on performance.* Biological Psychology, 2006. **73**(3): p. 242-252.

Woody, C. D., *Characterization of an Adaptive Filter for the Analysis of Variable Latency Neuroelectric Signals,* Medical and Biological Engineering, 1967. 5:p. 539-553.

DISTRIBUTION

1	MS0359	D. Chavez, LDRD Office	1911
1	MS0899	Technical Library	9536 (electronic copy)
1	MS1327	Laura Matzen	01463
1	MS1327	Michael Haass	01463
1	MS1327	Phil Bennett	01463

Sandia National Laboratories