Neural correlates of false memory disqualification by true recollection of feedback

Taylor M. Joerger[a] and Jennifer A. Mangels[a,b]

[a]Psychology Department, Columbia University and [b]Psychology Department, Baruch College, City University of New York, New York, USA

Correspondence to Dr Jennifer Mangels, Baruch College, CUNY, NY 10010, USA
Tel: + 646 319 4370; fax: + 646 312 3781; e-mail: jenimangels@gmail.com

Received 19 July 2008; accepted 20 August 2008

DOI: 10.1097/WNR.0b013e328315cd1c

Event-related potentials associated with disqualifying false memories were recorded in a novel false memory paradigm in which participants were given feedback during an initial recognition test, followed by a surprise retest where true recollection of feedback could be used to disqualify previous errors. Two spatiotemporally distinct components emerged: a parietal left-lateralized positivity indexing the recollection of feedback (500–900 ms), which was subsequently joined by a bilateral frontocentral positivity (700–900 ms) associated with rejection of the erroneous response and/or switching to the correct response. This latter effect seems to be distinct from the more anterior and later right frontal positivity typically associated with postretrieval monitoring. *NeuroReport* 19:1695–1698 © 2008 Wolters Kluwer Health | Lippincott Williams & Wilkins.

Keywords: event-related potentials, false memory, feedback, frontal lobe, late positive component, recollection, retrieval, switching

Introduction

In past years, much research has been dedicated to understanding processes underlying memory retrieval errors [1,2], particularly those coupled with strong subjective experiences of veridicality (i.e. false memories). Cognitive neuroscience has contributed to this effort by comparing and contrasting neural mechanisms underlying the retrieval of false and true memories [3,4]. Building on these findings, behavioral research has revealed successful strategies for reducing the prevalence of false memories, with a particular emphasis on those involving retrieval monitoring [5]. Yet, relatively little research has investigated the neural correlates of their successful implementation.

In the Deese–Roediger–McDermott (DRM) paradigm [6], known for reliably eliciting high rates of false alarms, participants study words with relatively strong backward associative strength to a nonpresented theme word. When the theme word and/or other highly related distracters are later presented in an old/new recognition test, participants are likely to misidentify these items as 'old.' Two categories of retrieval monitoring strategies have been proposed to remediate these false alarms: diagnostic and disqualifying monitoring [7]. In diagnostic monitoring, participants identify memories as false if source details from the study phase cannot be retrieved, whereas disqualifying monitoring uses recollection of true events to reject the possibility that a false event occurred (i.e. 'recall-to-reject'). This study focuses on disqualification, which has received less attention regarding its neural correlates than diagnostic monitoring [8,9]. Here, we created a novel modification of the traditional DRM paradigm in which immediate veridical feedback (FB) is provided after each old/new decision. After initial test, participants take a surprise retest on the same items, providing them an opportunity to use FB recollection to avoid repeating any earlier false alarms.

Effective error remediation requires that participants not only recollect the FB associated with a particular word but also retrieve the erroneous response with which the FB was associated (e.g. 'old') to switch their response (e.g. from 'old' to 'new'). We used event-related potentials (ERPs) to monitor the timecourse and scalp topography of these processes. Although multiple ERP components are elicited during retrieval [10,11], disqualification is hypothesized to require explicit recollection of information [12]. Thus, we focused on the extent to which memory for FB elicited a late positive component (LPC), as this centroparietal positivity is often considered to be an index of recollection [13]. We also examined activity over frontal sites, which has been ascribed to various controlled retrieval processes including source retrieval [14], postretrieval monitoring [11], and response inhibition/switching [15], all of which may be important in using FB recollection to disqualify false memories.

Participants and methods

Written informed consent for a protocol approved by the Human Subjects Committee of Columbia University was obtained from 25 Columbia University undergraduates (12 female; age: M=19.8 years, SD=1.2). All participants were right-handed, native English speakers with normal psychological and neurological function. Participants were compensated at a rate of \$10/h with a \$10 bonus for completion. Five participants were excluded for poor behavioral performance ($d'<0.25$); four were excluded for low trial counts in critical ERP conditions.

The experiment was divided into three segments: study, initial test, and retest. Electroencephalogram (EEG) was recorded during the test phases, which took place approximately 24 h after the study phase.

At study, participants saw 28 lists of 15 words each centrally presented on a computer screen for 2 s each followed by a 500 ms fixation cross. For 10 s between each list, participants counted backwards by 3s from a three-digit number. Lists were derived from Roediger *et al.* [16] and contained words that were semantically related to some degree to a nonpresented theme word. Participants were told that they would be tested 24 h later, but they were not told about the additional retest. At study, they were prompted to think about the associations between words, which typically increases false alarm rates. Ordering of lists and words within lists was randomized.

During the initial test, participants made old/new and confidence decisions on the middle seven words from each studied list and seven nonstudied words for a total of 392 words. Participants saw each word alone for 2 s then were cued to make an old/new judgment while the word remained on the screen. They were told to respond 'old' only to items presented in the previous day's study phase. Participants received 1 s of accuracy FB, flanked by a 500 ms interstimulus interval. Negative FB was red and accompanied by a low tone. Positive FB was green and accompanied by a high tone.

In the surprise retest, participants were shown the same 14 words from the initial test phase, and an additional six retest-unique distracters for each list, for a total of 560 words. The six nontheme distracter words used at initial test and retest were randomly selected from a pool of 12 list-related words such that any given word was equally likely to be a distracter at initial or retest. The old/new test format was the same as in the initial test. They were reminded that any word not seen during the study phase was 'new', as from the initial test. Participants were then shown the following legend and asked to make a combined confidence/FB decision: 1=HC WFB, 2=HC NFB, 3=LC WFB, and 4=LC NFB. HC referred to high confidence, LC to low confidence, and 'WFB' referred to situations when initial test FB was used to make the response decision, and 'NFB' when it was not. No FB was provided at retest.

Continuous EEG was recorded from 64 sintered Ag/AgCl electrodes, digitized at 500 Hz, with a bandpass of 0.15–100 Hz. Impedance was kept below 11 kΩ. EEG was initially referenced to Cz then converted to an average reference off-line. We compensated for blinks and other ocular artifacts using BESA 5.1.8 (MEGIS Software, Germany). Off-line, EEG was cut into epochs time locked to retest word presentation, from −100 to 1000 ms poststimulus. Later, time windows were not analyzed, as visual inspection indicated no clear distinctions across conditions, perhaps because of greater artifact contamination related to difficulty in sustaining attention for long periods of time during the retest.

ERPs were analyzed from 500–700 ms to 700–900 ms. Separate analyses of variance (ANOVAs) were conducted across frontocentral (F3/4, FC1/3, FC2/FC4), central (C1/3, C2/4), and centroparietal (CP1/3, CP2/4, P3/4) regions. These analyses included the factors of hemisphere and electrode, as well as relevant conditions. Electrode effects were reported only if they interacted with condition. Greenhouse–Geisser corrections were made for violations of sphericity, and significant effects at the P value of less than 0.05 level were pursued with post-hoc t-tests.

Results

First, we focused on the prevalence of feedback retrieval and the neural correlates of this process. Thus, we examined items repeated across tests, and included only high confidence responses to eliminate guesses. A behavioral ANOVA assessing the relationship between retest accuracy (correct, incorrect) × FB retrieval (with, without) × item (target, distracter), revealed a three-way interaction, $F(1, 15)=5.9$, $P<0.03$. Correct retest responses (hits and correct rejections) and false alarms were primarily made in association with FB retrieval, whereas misses, which were rare overall, were made equally often with or without FB (misses: M1/417.6%, SE1/4.80; false alarms: M1/432.3%, SE1/4.80).

As illustrated in the top panels of Fig. 1, analysis of the corresponding ERPs (misses were not included) demonstrated a clear effect of recollecting FB from 500 to 700 ms, spanning frontocentral to centroparietal regions. For each of the three regions analyzed, significant interactions between FB and electrode (all $F>5.0$, all $P<0.05$) indicated that these effects were strongest near the midline. Similar results were found from 700 to 900 ms, except that they were focused primarily on central [FB: $F(1,15)=7.0$, $P<0.02$; FB × electrode: $F(1,15)=5.7$, $P<.04$] and centroparietal regions [FB: $F(1,15)=12.4$, $P<0.005$; FB × electrode: $F(1.4,21.6)=8.8$, $P<0.005$]; FB effects at frontocentral sites only reached marginal levels of significance ($P=0.08$). No significant effects of response (hit, false alarm, correct rejection) at any region or time period were observed. Although central and centroparietal waveforms were left lateralized (hemisphere: $F>8.0$, $P<0.03$), this asymmetry did not interact with condition.

We then examined how recollection of FB influenced the likelihood that responses would switch or stay at retest. Given the high overall hit rate to targets at both first test ($M=67.6\%$, SE=0.69) and retest ($M=68.8\%$, SE=0.78), only distracters were analyzed. We first assessed the probability of making a high confidence correct rejection at retest as a function of whether the initial test response had been a false alarm (switch) or a correct rejection (stay). A 2 (FB) × 2 (switch/stay) ANOVA revealed an overall effect of 'stay' responses being more common than 'switch' responses, $F(1,15)=35.0$, $P<0.001$. Importantly, this was qualified by a significant FB × response interaction, $F(1,15)=5.2$, $P<0.04$, such that when participants switched (i.e. corrected initial false alarms) they were more likely to use FB, whereas when they maintained correct rejections across tests, they were more likely not to use FB. A parallel analysis of high confidence false alarms at retest found that maintaining a false alarm across tests was also more prevalent than switching from a correct rejection to a false alarm, $F(1,15)=6.1$, $P<0.03$, but FB did not differentially support these switch versus stay responses.

In the bottom panels of Fig. 1 we contrasted ERPs associated with correct rejection or false alarm maintenance with those elicited by successful switching from a false alarm to a correct rejection. Given that error correction rarely happened without FB retrieval, only FB responses were considered [1]. A series of ANOVAs isolated significant condition effects in the later time period (700–900 ms), at

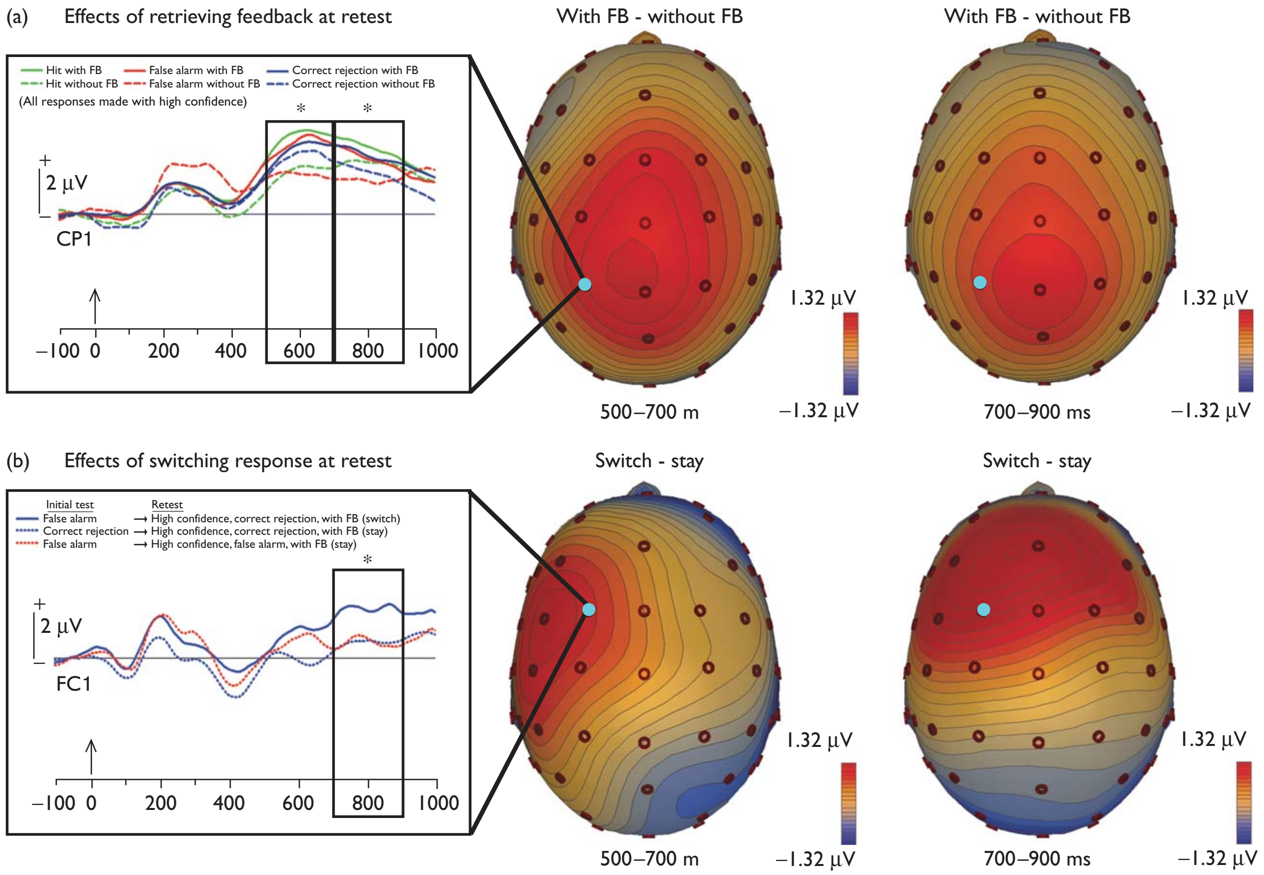

Fig. 1 (a) Effects of retrieving feedback at retest. Left side: grand mean waveforms (smoothed at 15 Hz), shown at a representative electrode. Right side: scalp topography highlighting differences between retrieval with and without feedback (FB) (collapsed over response). (b) Effects of switching response at retest. Left side: grand mean waveforms (smoothed at 15 Hz), shown at a representative electrode. Right side: scalp topography highlighting differences between switch condition and average of the two stay conditions.

frontocentral sites only, $F(1.3,19.8)=6.0$, $P<0.02$. Post-hoc comparisons indicated that frontal activity was greater for the successful disqualification of an initial false alarm, compared with when FB was recollected but participants were unsuccessful in correcting the initial error (false alarm-stay) or feedback was simply used to maintain a correct rejection across tests (correct rejection-stay). At central and centroparietal sites, activity was left-lateralized overall ($F>5.0$, all $P<0.05$), but no significant interactions between hemisphere and condition emerged.

Discussion

This study examined the processes engaged when participants successfully used recollection of corrective FB to disqualify false memories made on an initial test. Although the initial retrieval of FB was accompanied by centroparietal positivity across a broad interval (500–900 ms), implementation of FB to remediate errors involved a more frontally distributed positivity, maximal in the later part of this time window (700–900 ms). The spatiotemporal distribution of the FB retrieval effect was similar to that of the typical LPC [13], supporting the view that disqualification involves initial recollection of information that excludes the possibility of a false alarm. The later frontal activity seemed to index the use of this information to reject the false memory, which in this paradigm involved inhibition of the earlier response and/or switching to the correct response.

To our knowledge, in addition to being the first to explore the neural correlates of disqualification, the use of explicit FB is a novel experimental approach to reducing false recognition. Prior disqualification studies have focused either on source-based exclusion rules [17], an exhaustive recall-to-reject strategy effective only for very short lists (three items) [7], or multiple study-test trials without FB [18], which require participants to keep track of earlier test responses during later study trials to eventually 'tag' distracters as nonstudied. We expected that FB would be an effective method for rapid disqualification of false memories given our previous studies showing its utility in correcting retrieval errors in the context of a general knowledge retrieval task [19]. As participants in that paradigm still recall the initial erroneous response even after correctly rejecting it [20], it is likely that FB is disqualifying rather than overwriting the error.

Our ERP analysis established a relationship between the LPC and successful retrieval of a salient FB event, thereby linking this waveform to the recollection of critical information that could be used to 'reject' the false memory. Thus, it might seem surprising that this parietal waveform did not demonstrate a typical 'old/new' effect, whereby correct 'old' items show an enhanced ERP compared with correct 'new' items. Unlike earlier studies, however, the distracters we analyzed at retest were not truly novel because they had been presented once at initial test.

Participants automatically encode test distracters, even when not expecting an upcoming retest [21]. Thus, correct rejections at retest may have been accompanied with recollection of the initial test experience, similar to that of hits and false alarms, as is supported by similar levels of FB retrieval across these response categories. Yet, participants were still able to reliably discriminate twice-repeated studied items from these once-repeated distracters (d' for high confidence responses: M=1.04, SE=0.07), thus hits may have been distinctive on the basis of some additional study-phase information. The quantity of this information, however, may have been small relative to the recollection of FB [2], especially given that it would have had 24 h to decay, whereas the initial test and retest occurred in the same session. Thus, a significant additive old/new effect might have emerged with additional participants, or alternatively, large old/new effects may have been present that were outside the spatiotemporal scope of our analyses.

The disqualification of false alarms was also associated with bilateral frontal activity that emerged from 700 to 900 ms. Neuroimaging studies have often identified equal or greater late-onset anterolateral frontal activity during false compared with true recognition [3,22,23], suggesting that this region is engaged more by the increased effort or uncertainty associated with postretrieval monitoring than by the successful use of this monitoring to reject lures. Indeed, less frontal activity is found when false alarms can be easily avoided on the basis of highly accessible diagnostic information [8,9]. Given that we only compared high confidence responses, it is unlikely that this activity reflected uncertainty. Our frontal activity, which was more bilateral and posterior than waveforms typically associated with postretrieval monitoring, was not present in the false alarm 'stay' condition, also suggesting that it was unrelated to any additional monitoring or evaluation required by false alarms. Its absence in the correct rejection 'stay' condition further indicates that it was unrelated to any additional strategic demands of retrieving the correct FB-response association.

Rather, we propose that this frontal activity indexes processes whereby recollection of the FB is used to switch from a false alarm to a correct rejection. Frontal activity, encompassing medial and/or lateral regions, is commonly found in various task-switching paradigms [15,24], many of which require not only task-set reconfiguration, but also inhibition of previous stimulus–response associations [15,25]. Thus, the present results suggest that when information is retrieved from long-term memory for the purpose of 'rejecting' an erroneous response, lateralized and sustained postretrieval frontal activity is engaged. However, the latency of this effect was slightly earlier than the traditional postretrieval frontal effect (1000–1500 ms), and therefore, we cannot rule out the possibility that additional monitoring may have occurred post-switching.

Conclusion

Disqualification could be dissociated into two spatiotemporally distinct components: a parietal LPC associated with initial recollection of the disqualifying FB and a subsequent frontocentral positivity associated with inhibiting the incorrect response and/or switching to the correct response.

Acknowledgements

The authors thank Brian Maniscalco for invaluable assistance in participant testing and programming of analysis routines. Brady Butterfield and Matt Greene contributed to programming the paradigm. This research was supported by an NIH grant R21MH066129 to J.A. Mangels.

References

1. Loftus EF, Polage DC. Repressed memories. When are they real? How are they false? *Psychiatry Clin N Am* 1999; **22**:61–70.
2. Lampinen J, Neuschatz J, Payne D. Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Curr Psychol* 1998; **16**:181–224.
3. Schacter DL, Slotnick SD. The cognitive neuroscience of memory distortion. *Neuron* 2004; **44**:149–160.
4. Schacter DL, Norman KA, Koutstaal W. The cognitive neuroscience of constructive memory. *Annu Rev Psychol* 1998; **49**:289–318.
5. Gallo D. *Associative illusions of memory: false memory research in DRM and related tasks*. Psychology Press: New York; 2006.
6. Roediger HL, McDermott KB. Creating false memories: remembering words not presented in lists. *J Exp Psychol: Learning, Memory, Cognition* 1995; **21**:803–814.
7. Gallo DA. Using recall to reduce false recognition: diagnostic and disqualifying monitoring. *J Exp Psychol Learn Mem Cogn* 2004; **30**:120–128.
8. Gallo DA, Kensinger EA, Schacter DL. Prefrontal activity and diagnostic monitoring of memory retrieval: FMRI of the criterial recollection task. *J Cogn Neurosci* 2006; **18**:135–148.
9. Budson AE, Droller DB, Dodson CS, *et al*. Electrophysiological dissociation of picture versus word encoding: the distinctiveness heuristic as a retrieval orientation. *J Cogn Neurosci* 2005; **17**:1181–1193.
10. Friedman D, Johnson R Jr. Event-related potential (ERP) studies of memory encoding and retrieval: a selective review. *Microsc Res Tech* 2000; **51**:6–28.
11. Rugg MD. Retrieval processing in human memory: electrophysiological and fMRI evidence. The Cognitive Neurosciences. 3rd ed. In: Gazzaniga MS, editor. Cambridge: MIT Press; 2004.
12. Gallo DA, Bell DM, Beier JS, Schacter DL. Two types of recollection-based monitoring in younger and older adults: Recall-to-reject and the distinctiveness heuristic. *Memory* 2006; **14**:730–741.
13. Rugg MD, Curran T. Event-related potentials and recognition memory. *Trends Cogn Sci* 2007; **11**:251–257.
14. Kuo TY, Van Petten C. Perceptual difficulty in source memory encoding and retrieval: Prefrontal versus parietal electrical brain activity. *Neuropsychologia* 2008; **46**:2243–2257.
15. Monsell S. Task switching. *Trends Cogn Sci* 2003; **7**:134–140.
16. Roediger HL III, Watson JM, McDermott KB, Gallo DA. Factors that determine false recall: A multiple regression analysis. *Psychon Bull Rev* 2001; **8**:385–407.
17. Dodhia RM, Metcalfe J. False memories and source monitoring. *Cogn Neuropsychol* 1999; **16**:489–508.
18. McDermott K. The persistence of false memories in list recall. *J Mem Lang* 1996; **35**:212–240.
19. Butterfield B, Mangels JA. Neural correlates of error detection and correction in a semantic retrieval task. *Brain Res Cogn Brain Res* 2003; **17**:793–817.
20. Butterfield B, Metcalfe J. Errors committed with high confidence are hypercorrected. *J Exp Psychol Learn Mem Cogn* 2001; **27**:1491–1494.
21. Stark CE, Okado Y. Making memories without trying: medial temporal lobe activity associated with incidental memory formation during recognition. *J Neurosci* 2003; **23**:6748–6753.
22. Goldmann RE, Sullivan AL, Droller DB, *et al*. Late frontal brain potentials distinguish true and false recognition. *Neuroreport* 2003; **14**:1717–1720.
23. Nessler D, Mecklinger A, Penney TB. Event related brain potentials and illusory memories: the effects of differential encoding. *Brain Res Cogn Brain Res* 2001; **10**:283–301.
24. Wager TD, Jonides J, Reading S. Neuroimaging studies of shifting attention: a meta-analysis. *Neuroimage* 2004; **22**:1679–1693.
25. Aron AR, Monsell S, Sahakian BJ, Robbins TW. A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. *Brain* 2004; **127** (Pt 7):1561–1573.