Available online at www.sciencedirect.com
SCIENCE DIRECT
Cognitive Brain Research 17 (2003) 793–817
COGNITIVE BRAIN RESEARCH
www.elsevier.com/locate/cogbrainres

Research report

# Neural correlates of error detection and correction in a semantic retrieval task

Brady Butterfield*, Jennifer A. Mangels*

*Department of Psychology, Columbia University, 1190 Amsterdam Ave., New York, NY 10027, USA*

Accepted 13 August 2003

**Abstract**

Event-related potentials (ERPs) were used to investigate the cognitive and neural substrates of immediate and 1-week delayed error correction in a semantic retrieval task. In particular, we pursued the basis for the 'hypercorrection' effect, the finding that erroneous responses endorsed as correct with high confidence are more likely than low-confidence errors to be corrected at retest. Presentation of negative, but not positive feedback about the accuracy of one's response elicited a fronto-central negativity, similar to the ERN, which was somewhat sensitive to the degree to which negative feedback violated expectation. A fronto-central positivity, similar to the novelty-P3/P3a, more generally indexed detection of a metamemory error, given that it was larger in conditions of high metamemory mismatch than in conditions of low metamemory mismatch, irrespective of absolute task accuracy. For errors, amplitude of the fronto-central positivity, but not the preceding negativity, was correlated with correction on an immediate retest. Thus, to the extent that the fronto-central positivity indexes an orienting response, this response appears to facilitate initial encoding processes, but does not play a key role in memory consolidation. In contrast, a broad, inferior-temporal negativity occurring 300–600 ms after presentation of the correct answer was sensitive to subsequent memory performance at both immediate and delayed retests, but only for answers containing familiar semantic information. This negativity may reflect processes involved in the formation of an association between the question and pre-existing semantic information.


*Theme:* Neural basis of behaviour

*Topic:* Cognition

*Keywords:* ERN; Ne; Pe; P3; P3a; Error; Metamemory; Mismatch; Encoding

## 1. Introduction

### 1.1. Overview

Errors are common in all realms of human cognition. Within the domain of memory, errors include failures of information retrieval (misses), as well as the erroneous retrieval and endorsement of false information (intrusions and false alarms). Over the past 20 years, much research has been dedicated to characterizing the conditions under which these errors, particularly false alarms, are most likely to occur (for reviews, see Refs. [52,56]). Only recently has the focus turned to how these errors can be avoided or suppressed [16,30]. In the present studies, we examine the cognitive and neural basis underlying how errors in retrieval from semantic memory, once made, can be subsequently corrected. In particular, we focus on the relationship between the metamemory one has for the predicted accuracy of their answer, reflected by response confidence, and the likelihood that the corrective feedback will be successful. For example, after one confidently states that the capital of Canada is Toronto, how can this misinformation be replaced with the correct response of Ottawa?

### 1.2. Hypercorrection effect

Many models of memory would predict that an erroneous response to this type of question would be more

*Correspondence can be addressed to either author. Tel.: +1-212-854-3243; fax: +1-212-854-3609.
*E-mail addresses:* bab67@columbia.edu (B. Butterfield), mangels@psych.columbia.edu (J.A. Mangels).

0926-6410/03/$ – see front matter 
doi:10.1016/S0926-6410(03)00203-9

difficult to correct if a subject had initially felt very confident in the accuracy of his or her answer (e.g., Refs. [1,35,59]). Given that confidence ratings are at least partially based on retrieval fluency [3,45], and retrieval fluency is at least partly related to the strength of association between the question and answer, we can infer that the question–answer association would be stronger for the incorrect answers that had been endorsed with high confidence compared to incorrect answers that had been endorsed with low confidence. As a result, one could predict that it should be more difficult to displace these high-confidence errors with the correct information.

In contrast to these predictions, a recent study by Butterfield and Metcalfe [6] demonstrated that corrective feedback is actually more effective in remediating errors when those errors are initially endorsed with high, rather than low, confidence. In that study, participants first provided responses to trivia questions, then rated their confidence in the accuracy of that response. They were then given immediate feedback about the accuracy of their response and informed of the correct response if they were in error. Later, at a surprise retest, they were instructed to list the first three responses that came to mind for a subset of both initially correct and incorrect questions and indicated the response that they now believed was correct. When a subject had been highly confident in the accuracy of an erroneous response at the first test, the subject was actually more, rather than less, likely to provide the objectively correct answer in this list and endorse that answer as correct at retest. The learning benefit associated with making a more egregious error was termed 'hypercorrection.'

The erroneous information had not been forgotten, however. High-confidence errors were more likely than low-confidence errors to be given as one of the three retest responses, indicating that the erroneous information was still readily accessible. Yet, there is little evidence that hypercorrection was simply the result of direct mediation, in which participants learn the correct answer by associating it with the incorrect answer. The presence of the error did not predict the presence of the correct answer in the list of three responses. Next, we present two additional hypotheses to explain the superior learning arising from high-confidence errors: 'metamemory mismatch' and 'domain familiarity.' Note that these explanations are not mutually exclusive.

#### *1.2.1. Metamemory mismatch*

Metamemory refers to beliefs concerning one's memory capabilities [63], and can reflect self-assessment of the results of a specific retrieval attempt, as well as an evaluation of one's overall memory strengths, weaknesses and available strategies. In the present studies, a confidence judgment is used to indicate a subject's metamemorial expectations about the accuracy of his/her response, which is then validated, or not, by the feedback. A judgment of high confidence means that the subject predicts that their answer is correct, whereas a judgment of low confidence means that the subject predicts that his/her answer is incorrect. Although all error feedback signals a conflict between a subject's response and task demands, in the case of high-confidence errors there is an additional conflict between these metamemorial expectations and the actual outcome—a conflict which we term 'metamemory mismatch.' This additional conflict may result in a stronger or more effective trigger to control processes that are capable of initiating remedial actions, such as increased attention to the correct answer and inhibition of the incorrect answer (e.g., Ref. [5]). Thus, the metamemory mismatch bears similarity to the influential animal model of associative learning formulated by Rescorla and Wagner [74], in which learning is fastest when deviation between actual outcome and the subjective likelihood of that outcome is greatest.

The attentional effect of this conflict receives further support from a series of two dual-task experiments by Butterfield and Metcalfe [7]. They added a tone-detection task to a general information question task highly similar to the one employed in the present experiments. It was found that, when quiet tones were played during the presentation of feedback, participants were less likely to report hearing a tone presented during metamemory-mismatch feedback (i.e., feedback to high-confidence errors or low-confidence corrects) than during metamemory-match feedback (low-confidence errors or high-confidence corrects). This result was interpreted to be due to the capture of attention by the high metamemory mismatch, leaving less attention available for tone detection. For errors, failure to detect a tone was also associated with improved memory for the feedback at retest.

#### *1.2.2. Domain familiarity*

An alternative hypothesis for the hypercorrection effect focuses on the semantic landscape surrounding high- and low-confidence responses. Specifically, confidence ratings may be influenced by how much domain-relevant information a question activates, and/or how much a participant feels s/he knows about the topic of the question. In the case of errors, familiarity with a given question's domain (e.g., Canadian geography for the question above) might increase the likelihood that the erroneous response is endorsed with high confidence. High domain familiarity increases the likelihood that the correct information is already stored in semantic memory, even if it were not associated with the question strongly enough to be given as the response at first test. Learning to associate pre-existing information in semantic memory would be easier than encoding completely novel information. For example, according to Thorndike's [85] conceptualization of learning, associating a response one has never heard of before with a question would require two stages of learning: first learning the response as an entity in and of itself and then

associating the response with the question. If one were already familiar with the correct response, however, only the second, easier stage of learning would be required.

### 1.3. Electrophysiological correlates

In the present studies, we explore the neural correlates associated with these proposed components of error correction by measuring event-related potentials (ERPs) elicited when subjects are given feedback regarding their response to trivia questions. This feedback informed subjects not only about the accuracy of their response, but also the correct answer, either simultaneously (Experiment 1) or successively (Experiment 2). ERPs provide a means of non-invasively tracking, with millisecond-level resolution, fluctuations in the activity of neuronal populations that are consistently time-locked to the onset of stimulus (or response) processing. As such, ERPs have been a method of choice in studying the sequence of rapid neurocognitive processes that characterize the detection of errors and conflict. In addition, ERPs can be selectively averaged according to an individual's performance, a feature that has been exploited in many studies investigating processes associated with successful memory encoding. To isolate such processes, the electrophysiological activity recorded during a study phase is selectively averaged according to the retrieval outcome. ERPs to items later retrieved are then compared to ERPs of items later forgotten, and any differences between them are interpreted as indicating specific aspects of stimulus processing that are predictive of encoding success. We use this method of selective averaging to examine what aspects of the neurocognitive response to the accuracy and corrective (correct answer) feedback are related to successful correction of errors when retested immediately (Experiment 1) and after a 1-week delay (Experiments 1 and 2).

#### 1.3.1. Metamemory mismatch: ERPs associated with error and novelty detection

The metamemory mismatch hypothesis of hypercorrection focuses on the ability of the accuracy feedback to engage error detection processes. These error detection processes may serve to activate control processes designed to facilitate error correction, which for the purposes of this task, involves successful encoding of the correct answer. Thus, our investigation of the neural correlates associated with metamemory mismatch focused on ERPs previously shown to demonstrate sensitivity to the detection of error or distinctive events and have been linked to cognitive control.

The error-related negativity (ERN [32], also called the $N_E$ [22]) satisfies both of these criteria [5]. This frontocentrally maximal deflection appears to be elicited when a generic error evaluation system first detects a mismatch between one's response and the internalized goals of the task [10,39]. Whether the ERN is larger to the erroneous response itself or to feedback signaling response accuracy depends on the nature of the task's stimulus–response mapping. If the stimulus–response mapping is unambiguous, as is the case in many choice reaction time (RT) or go/no-go tasks, errors more likely result from premature responding rather than response uncertainty. Under these conditions, a large ERN is observed approximately 80–100 ms after the erroneous response (100–150 ms after the EMG onset) [23,32], but only a very small ERN is found to any subsequent feedback regarding response accuracy [39,66]. This is because errors can be determined during response execution by internal evaluation of an 'efferent response copy,' and, therefore, explicit performance feedback is redundant. In contrast, if the mapping of the correct response is less certain, as in the case of time estimation [61], the early stages of trial-and-error learning [39,66], or situations where the stimulus–response mapping is apparently random [39,66,75], little or no ERN is elicited by erroneous responses, but a large ERN is elicited when the error is signaled by accuracy feedback. In these situations, the subject is dependent on the feedback to determine whether an error has occurred. Such is the case in the present studies, in which subjects must wait for feedback to validate the accuracy of their answers to the trivia questions, regardless of whether they endorse their responses with low or high confidence. Only when subjects fail to provide *any* plausible answer (i.e., 'omit responses') is the outcome certain before feedback is presented. Thus, we would expect all errors to elicit a feedback-locked ERN, with the possible exception of omit responses.

Of particular interest is the extent to which an ERN elicited by error detection would be amplitude-modulated by the subject's expectations regarding performance outcome (i.e., metamemory), and, furthermore, whether any observed modulation would be related to the success of correcting the error at retest. The most compelling evidence for a relationship between subjective expectation and the ERN comes from a series of studies in which probability of stimulus–response mapping was manipulated [39]. Specifically, in a version of the probabilistic classification task used by Nieuwenhuis et al. [66], stimulus–response mappings were consistent 100, 80, or 50% of the time, and both response-locked and feedback-locked ERNs were measured. Critically, the amplitude of the ERN to the feedback was largest in the 80% 'invalid' condition—a condition in which a particular stimulus–response association that had been rewarded 80% of the time was not rewarded. Although confidence ratings were not taken after responses in this task, it is likely that subjects in the 80% condition would have approached the feedback fairly confident that their answer was correct. In the 50% consistent condition, however, it is likely that subjects would have difficulty generating strong expectations of any kind. Correspondingly, the 50% condition elicited a smaller feedback-locked ERN than the 80% invalid condition. The 100% consistent condition elicited

the smallest ERN, presumably because, in this condition, subjects would already be aware of their error by the time feedback was presented, a hypothesis supported by the large response-locked ERN in this condition. Together, these results support a recent general model of the ERN proposed by Holroyd and Coles [39], in which the ERN is hypothesized to reflect an error in reward prediction signaled by input from the mesencephalic dopamine system to the anterior cingulate cortex (ACC). In this model, larger discrepancies in predicted and actual reward elicit a larger negative-reinforcement learning signal by the mesencephalic dopamine system (see also Refs. [26,79]) and, in response, a larger ERN is generated by the ACC.

In the present studies, however, expectation is defined by a subject's metamemory—a conscious evaluation of the veridicality of his/her memory, rather than by stimulus–response probabilities. It is not clear whether the ERN would be equally sensitive to this type of expectation given that, in at least one study, the ERN was not modulated by subjective awareness. Specifically, Nieuwenhuis et al. [65], using an anti-saccade task, found that the amplitude of the response-locked ERN was uninfluenced by subjective awareness of an error (see also Ref. [57]). Both perceived and unperceived saccade errors generated equivalent ERNs, thus conscious evaluation processes such as metacognition might have little influence on the error monitoring processes indexed by this component. Conscious awareness of an error may be indexed, instead, by the amplitude of a positivity immediately following the ERN, termed the $P_E$, which was only observed for perceived saccade errors in that task. Although less work has been done to establish the functional significance of the $P_E$, others have noted that the longer latency of the $P_E$ is more consistent with a process that occurs either coincident with, or following, conscious response proprioception or stimulus evaluation [22,23,43,88]. Thus, to the extent that registration of metamemory mismatch relies on conscious evaluation of the error, confidence in the error may modulate the $P_E$ in addition to, or instead of, the ERN.

The $P_E$ has been defined primarily in relationship to commission a response error rather than perception of negative feedback, however some researchers have noted its similarity to a P3-like stimulus evaluation component [14,22,23,57]. In addition, previous studies of ERPs elicited by performance feedback have noted a vertex-maximal P3 that was larger for high-confidence errors than for low-confidence errors [15,40]. Although these studies used a relatively limited number of electrodes, the P3 in those studies bears similarity to the posterior P3b, or 'classic P3,' evoked by rare target stimuli in the auditory 'oddball' paradigm [51,70]. The extensively studied P3b component is thought to index processes associated with detection of a rare, task-relevant stimulus, such as stimulus categorization [51] and the updating of information in working memory [18]. Like rare target stimuli, high-confidence errors (metamemory mismatches) occur relatively infrequently (comprising only 5% of all errors in Ref. [6]), thus we might also expect categorization of feedback to these errors to modulate the amplitude of a posterior positive waveform similar to the P3b. More recently, however, a study that used motivational factors to increase the emotional salience of the error, such as social comparison and a monetary penalty for errors (but no gain for corrects), found that a $P_E$ peaking 280 ms after response onset had a more anterior distribution than the typical parietal P3b [57]. Spatiotemporally, this $P_E$ was more similar to the novelty-P3/P3a, a frontally-maximal waveform occurring just prior to the P3b that is also elicited by rare target stimuli, but unlike the P3b is also large following rare nontargets (not requiring a response) and other events that are novel in a given context (e.g., Refs. [11,13,36,47,49,81]).[1] This waveform is hypothesized to index detection of and/or involuntary switching of attention to a deviant event [27]. Given that in Butterfield and Metcalfe [6] high-confidence errors were not only rare, but also reported as more embarrassing and emotionally arousing than other types of errors, we may find that subjective expectation also modulates a feedback-locked positivity that has a more anterior distribution.

In order for metamemory mismatch to serve as a plausible explanation for the hypercorrection effect, however, the ERP component sensitive to this process must also be predictive of error correction. Although some studies have found the amplitude of the response-locked ERN in a choice RT task to be positively related to post-error slowing [32,57], a type of remedial action (see Ref. [73]), many others have not [31,33,61,65,78]. The relationship of the $P_E$ to error correction is also unclear. Nieuwenhuis et al. [65] found that post-error slowing only occurred after perceived errors, and concluded that the $P_E$, which also was only found for perceived errors, indexed error-monitoring processes that were associated with error correction. However, the ability to draw further predictions from this data to the present task is limited by the fact that nearly all these studies examined response-locked error potentials, rather than feedback-locked potentials, and examined these potentials in relation to immediate motor slowing rather than to long-term learning. The one study that examined the relationship of a feedback-locked error potential to subsequent error correction only measured the ERN, even though a $P_E$ was also apparent [61]. On the other hand, previous studies of learning in which the P3 was modulated by metamemory mismatch did not test whether the amplitude of this positivity was related to later

[1] Strictly speaking, the P3a refers to a frontally-maximal positivity elicited by rare, but expected stimuli in a traditional, two-stimulus oddball task, whereas the novelty-P3 refers to a positivity of similar latency and topography elicited by rare and unexpected stimuli inserted into a two-stimulus oddball task. Increasingly, these terms have been used interchangeably and recent research indicates that they are indistinguishable [80,82].

error correction [15,40]. However, a P3 with a posterior maximum and a longer latency has been related to long-term memory encoding in many other studies (Refs. [20,76], but see Ref. [58]). Thus, to the extent that error-related potentials are observed in the present studies, our results will provide further information about the relationship of error-related potentials to error correction in general.

#### *1.3.2. Domain familiarity: ERPs associated with semantic processing*

This hypothesis of hypercorrection suggests that the amount of knowledge an individual possesses on a particular topic influences both the level of confidence with which they will endorse incorrect answers to questions in that domain, and the likelihood they will find themselves familiar with the correct answer. Pre-existing familiarity with the correct answer means that subjects need only to strengthen the association between the question and answer in order to correct their error, rather than store novel information.

Given that obtaining a neural correlate of general domain familiarity would be difficult, we looked instead for a neural response that was modulated by the individual's familiarity with the correct answer, which was assessed for errors only, after presentation of the correct answer. In addition, because we predicted that familiarity would facilitate encoding, we were particularly attuned to the presence of ERPs that were sensitive to both familiarity and subsequent memory performance. Scalp-recorded ERP negative waves with latencies between 300 and 500 ms post-stimulus have been studied extensively in relation to lexical and semantic processing [53,54]. However, the most well studied of these waves—the classic N400 wave elicited over parietal regions by semantically incongruous stimuli [55]—does not appear to be predictive of later memory [64]. Moreover, for individuals who are familiar with a particular domain, familiar correct answers should be even more highly congruent with their preceding semantic context than unfamiliar answers, and would not be expected to elicit a large N400.

Therefore, rather than focusing on the classic N400, we turned our attention to a family of early negative waveforms, maximally recorded over left fronto-temporal sites, which have demonstrated positive relationships with both semantic processing and verbal memory [19,37,53,67,68]. For example, Nobre and McCarthy [67] found an early left fronto-temporal negativity (N316) time-locked to word processing that was enhanced when the preceding item was semantically related, suggesting that activity in this region may be modulated by the fluency of access to information in semantic memory. Fluency of semantic access is often used by subjects as an index of familiarity (e.g., Ref. [42]). A negative waveform similar in spatial distribution, but peaking slightly later (N340) also was found to predict subsequent episodic memory for words [58]. Based on its similarity to other waveforms associated with semantic processing, as well as cerebral blood flow demonstrating activation of left inferior frontal and left middle temporal cortex during retrieval from semantic memory (e.g., Refs. [28,90]), Mangels and colleagues interpreted the N340 as representing the initial, item-specific semantic processing of verbal information, a necessary condition of successful episodic encoding. Together, these results suggest that early negative-going waveforms in this region not only may provide an index of a subject's familiarity with the correct answer, but also the extent to which that familiarity facilitates the ability to produce that answer on a subsequent retest.

### *1.4. Summary*

Investigation of the ERP correlates of error detection and domain familiarity, and the relationship of these correlates to error correction were addressed across two experiments employing trivia questions from a wide range of topics. In both experiments, subjects provided an answer to the question and then indicated their level of confidence in that answer's accuracy. Subjects were then given feedback about their performance. In Experiment 1, this feedback simultaneously informed subjects about their accuracy and the correct answer. If the subject had made an error, we then assessed whether the subject was familiar with the correct answer. The success of error correction was based on subjects' performance on a surprise retest for items missed at the first test that was given a few minutes after completion of the first test. In Experiment 2, we presented the accuracy feedback and correct answer sequentially in order to better isolate the neural components associated with metamemory mismatch and familiarity. We also added a 1-week delayed retest in order to determine whether the relative contributions of these processes to error correction differed as a function of study-test delay.

## 2. Experiment 1

### *2.1. Method*

#### *2.1.1. Participants*

Twenty-five participants (14 females, mean age 21 years) were tested with ERPs. ERP data from five participants were lost or excluded due to computer problems, insufficient trial counts, or excessive noise. Participants were native speakers of English, had corrected-to-normal vision and had no history of neurological or psychological disorder. Prior to ERP testing, participants were screened for their knowledge of trivia using a pretest consisting of five easy, medium, and difficult questions that were not repeated during the actual test. They were excluded from ERP testing if they answered fewer than five or more than 12 questions correctly on this pretest. All participants gave

informed consent and were compensated at a rate of $10/h for their participation.

### 2.1.2. Materials

The stimuli were 220 trivia questions from a variety of knowledge domains. All questions had answers that were single words three to eight letters in length (e.g., Q: ''What poison did Socrates take at his execution?'' A: ''Hemlock''). These questions were adapted from the Nelson and Narens [62] norms, and were similar to those used in Butterfield and Metcalfe [6].

### 2.1.3. Procedure

*2.1.3.1. Trivia question task.* The experiment consisted of two phases, a test and a surprise retest. EEG was recorded during the first test phase only. Trivia questions were presented in the center of the computer screen and participants were given an unlimited amount of time to type a response on the computer keyboard. If participants were not certain about the answer, they were encouraged to make an educated guess. However, if they felt that they could not come up with even a remotely plausible answer, they were told to type 'xxx.' These responses were classified as 'omit' responses. For all responses except omits, participants then rated their confidence in their response on a scale ranging from 1 (sure incorrect) to 4 (not sure if correct or incorrect) to 7 (sure correct). Participants were encouraged to use the entire scale.

Immediately following the confidence rating or 'xxx' response, a central fixation point appeared for 1 s. Feedback was then provided for 2.5 s in the form of the correct answer presented in green if it matched the participants' response or in red if it did not. Subjects were instructed to avoid blinking or moving during the feedback period. The program used a letter-matching algorithm to score the response as a match (75/100 or greater), non-match (less than 70/100), or borderline (greater than 70/100 and less than 75/100; pilot work determined that these responses were sometimes incorrect responses, and other times badly misspelled correct responses). Feedback to 'xxx' (omit) responses was always presented in red. Feedback to responses of borderline accuracy was also presented in red, but these trials were not included in any analyses. Following corrective feedback (i.e., the correct answer presented in red), participants indicated their familiarity with the correct answer (1=familiar, 2=not familiar). We emphasized to participants that they should only give a rating of 'familiar' if they were familiar with that specific person, place, or thing that the question was referring to, rather than simply familiar with the word in general.

Participants were given short breaks after each block of 55 questions. At completion of the first test, the recording electrodes were removed and the subject was allowed to wash off the gel. Approximately 8 min after the first test, the participant then began the retest phase, which consisted of a surprise retest for those questions that were incorrectly answered at first test. The sequence of trial events in the retest phase was the same as during first test with the exception that familiarity was not assessed.

*2.1.3.2. ERP recording.* Continuous EEG was recorded during the first test only using a 64-electrode Quick-Cap (Compumedics Inc.). EEG was amplified with Neuroscan SYNAMPS, and digitized at a rate of 500 Hz with a bandpass of 0.15–50 Hz. Inter-electrode impedance was kept below 11 kΩ. Eye movements and blinks were recorded from electrodes 1 cm lateral to the outer canthi (LO1/LO2) and over the infra-orbital ridge (IO1/IO2). The EEG was initially referenced to Cz, and then converted to an average reference, as suggested by the recording and analysis guidelines outlined by the Society for Psychophysiological Research [71]. We corrected ocular artifacts using two to four BESA components derived from a series of horizontal, vertical and blink eye-movements recorded prior to the trivia test [4]. This correction procedure involved a reduction in resolution from 2 to 4 ms.

*2.1.3.3. ERP analysis.* Off-line, the EEG was cut into 2 s epochs that were time-locked to presentation of feedback and baseline corrected using the 100 ms pre-feedback interval. Out of a total of 4400 epochs (220×20 subjects), 43 (<1%) were rejected because of a behavioral response of borderline accuracy at either first test or retest. Of the remaining 4357 trials, 272 (<7%) were rejected due to excessive noise (i.e., activity greater than ±250 μV for 18 participants and greater than ±350 μV for two subjects who consistently exhibited eye movements exceeding ±250 μV).

To achieve sufficient trial counts for the ERP analyses we collapsed the seven-point confidence scale into low (1–3), medium (4), and high (5–7) confidence categories. All conditions had an average of at least 10 trials per participant, except for the low-confidence corrects (mean, 6.2 trials; range, 1–28 trials), medium-confidence corrects (mean, 9.0 trials; range, 2–17 trials), and medium-confidence errors (mean, 9.5 trials; range, 3–20 trials). Despite these low trial counts, we chose to evaluate medium confidence trials separately given that medium confidence trials are instances in which the subject believed an outcome of correct or incorrect to be equiprobable. Thus, it is likely that this condition was qualitatively different from either the high- or low-confidence conditions, in which subjects thought it was more probable that their answers were correct or incorrect, respectively.

Our ERP analyses focused on waveforms that have been modulated by error, conflict, or familiarity in previous studies and were consistent with the spatio-temporal distribution of prominent waveforms in the grand means (see Figs. 1, 2 and 4). These waveforms included a fronto-central negative deflection similar to the feedback-

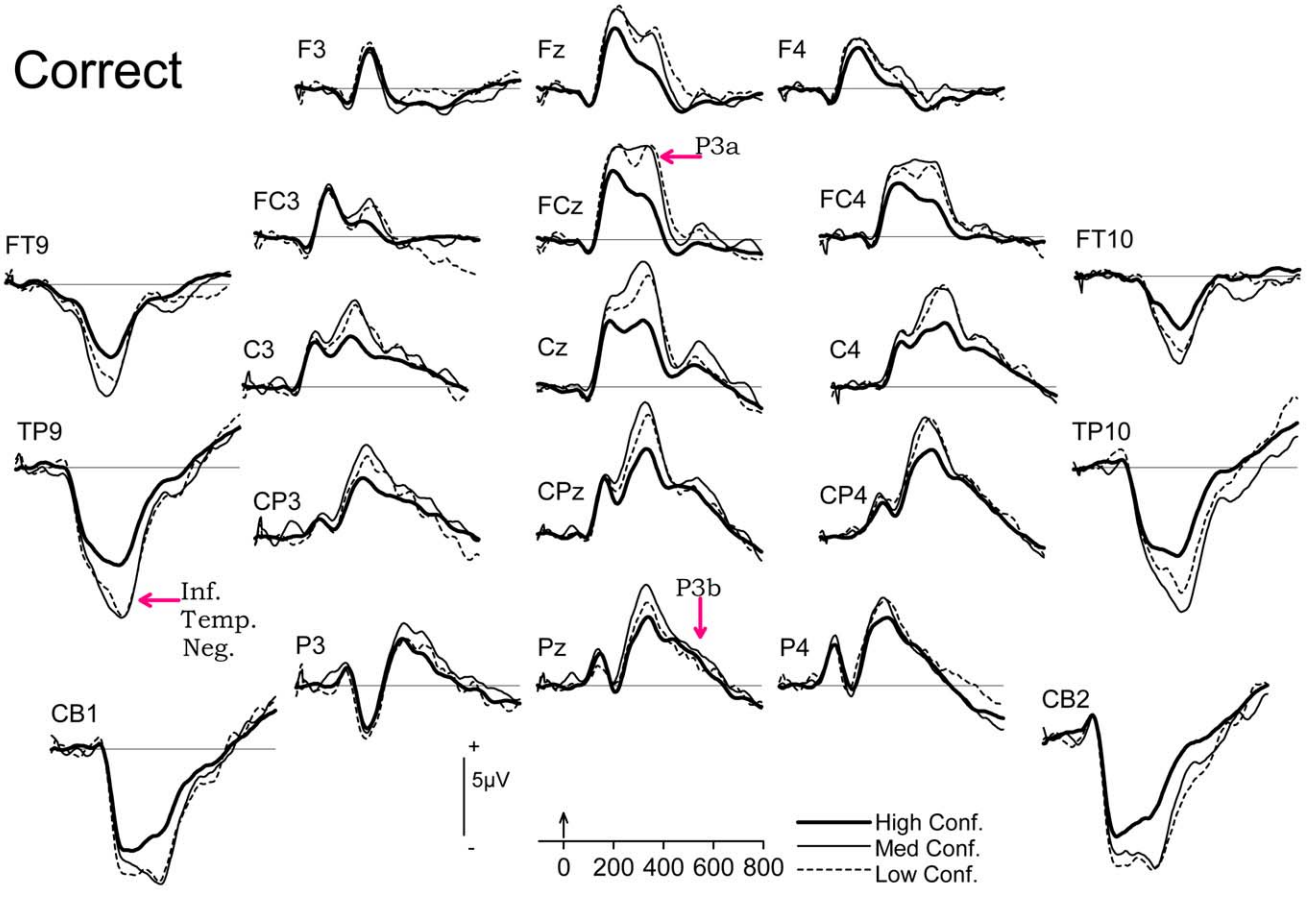


Fig. 1. Grand-mean waveforms at selected electrodes elicited by positive feedback (feedback following responses *correct* at first test) in Experiment 1, sorted by first test confidence. Data in this and all subsequent figures were low-pass filtered at 15 Hz.

elicited ERN [39,66] and a positive deflection immediately following that may be homologous to the novelty-P3/P3a [80]. Consistent with the typical distribution of the more anterior P3a, this positivity appeared to be attenuated at posterior sites. However, over posterior sites there was evidence of a distinct later and broader component that may be more homologous to the P3b [51]. Although we cannot be absolutely certain that the waveforms observed in this study are functionally equivalent to these positive components, which have been described mainly in the context of oddball stimulus-detection paradigms, for the purpose of clarity, in our description of the results we will refer to the fronto-central deflection following the ERN as the P3a and the later, broader and more posterior positivity as the P3b. We will consider more fully the proposed relationship of these components to the previous literature on the $P_E$, P3a and P3b in the discussion.

During the latency at which the P3a was maximal, and extending into the epoch at which the P3b was evident, a broader negative waveform could be observed at lateral sites that extended from fronto-temporal to more posterior regions along both hemispheres. Although this negativity may represent, in part, the inverse of the dipoles generating either or both P3 components, its differential sensitivity to the variables of interest in a previous study [58] suggests that it may also index activity from other sources. Therefore, we analyzed this waveform separately.

The waveforms of interest were preceded by a series of positive and negative deflections (i.e., P1, N1, P2) that represent the early stages of stimulus evaluation. We will not discuss these waveforms further except to note that the peak latency of the P2 was not found to be affected significantly by accuracy, confidence or their interaction (all $P>0.13$), supporting the view that the presence of the ERN in negative feedback trials is not simply an artifact of a latency shift in the positive deflections that flanked it.

Given that the ERN and P3a were present as consecutive deflections with a narrow temporal window, we analyzed these components by taking the mean amplitude of a 50 ms window centered on the peak latency of each deflection at FCz, where these deflections appeared to be maximal. Rather than simply taking the peak latency from the grand mean, however, we first determined whether there were any significant latency differences as a function of accuracy and/or confidence by measuring the peak of these two deflections at FCz in the six relevant conditions for each subject. There were no significant latency differences (all $P>0.12$) so the conditions were averaged to calculate

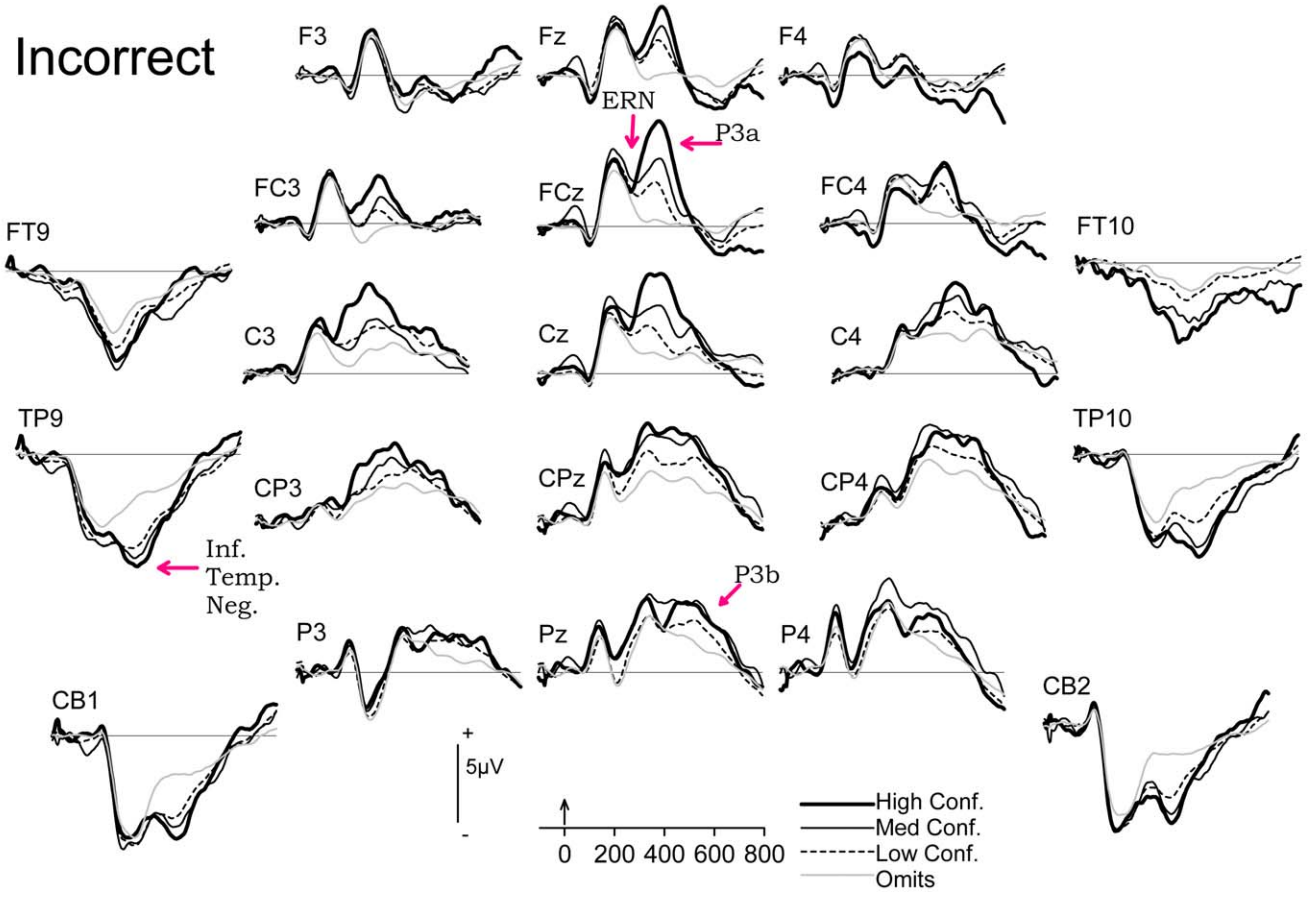


Fig. 2. Grand-mean waveforms at selected electrodes elicited by negative feedback (feedback following responses *incorrect* at first test) in Experiment 1, sorted by first test confidence.

mean peak latency for each of these waveforms. The mean peak latencies of the ERN and P3a were 276 and 352 ms, respectively, from which the 50 ms windows for calculating mean amplitude were centered. Because the P3a also has been described as having functionally dissociable frontal and posterior aspects in 'novelty oddball' paradigms [29] (for review, see Ref. [27]), we compared amplitudes at FCz and Pz measured during the P3a time window. These two electrodes were selected because they appeared to provide maximal distinction (minimal overlap) between frontal and posterior aspects. We also compared amplitudes at FCz and Pz in our analysis of the later P3b, which because it was broader in morphology, was measured as the mean amplitude from 450 to 650 ms. The mean amplitude of the inferior temporal negativity was analyzed in a broad window consistent with the morphology of this waveform (300–600 ms) at three pairs of electrodes running anterior–posterior along the inferior scalp (FT9/FT10, TP9/TP10, Cb1/Cb2).

All effects were analyzed with ANOVAs using the mean amplitude within the time windows specified above as the dependent variable. Greenhouse–Geiser corrections [41] were applied where appropriate. Epsilon values will be reported alongside uncorrected degrees of freedom. Significant main effects and interactions were followed by Tukey's HSD post-hoc comparisons and only significant comparisons are reported. The alpha level for all analyses was 0.05.

### 2.2. Results

#### 2.2.1. Behavioral results

Subjects answered an average of less than half of the questions correctly at first test (M=0.42, S.D.=0.13), but were able to correct about half of these errors at retest (M=0.52, S.D.=0.16). To assess whether subjects were hyper-correcting those questions initially answered incorrectly with high confidence, as well as the relationship of familiarity with the correct answer to retest correction, a series of within-subjects ANOVAs were computed. These ANOVAs are based on probabilities of responses being endorsed with each confidence level, correct at first test and retest, and unfamiliar or familiar, as shown in Table 1. First, we found that participants' confidence ratings were reliable indicators of first-test accuracy, $F(2,38)=195.1$, $P<0.001$. Post-hoc comparisons indicated that high-confidence responses were more likely to be correct at first test than low or medium confidence responses. There was also

Table 1
Conditional probabilities (with S.E.M.) of responses in Experiment 1

| *P*(response confidence) | | Given resp. conf. | *P*(correct at first test) | Given incorrect at first test and response confidence | *P*(correct at retest) | *P*(correct answer was familiar) |
|---|---|---|---|---|---|---|
| Omit | 0.35 (0.03) | Omit | – (–) | Omit | 0.41 (0.04) | 0.51 (0.03) |
| Low | 0.17 (0.02) | Low | 0.19 (0.03) | Low | 0.63 (0.04) | 0.68 (0.04) |
| Med | 0.09 (0.008) | Med | 0.48 (0.04) | Med | 0.65 (0.05) | 0.77 (0.04) |
| High | 0.40 (0.03) | High | 0.87 (0.01) | High | 0.79 (0.04) | 0.82 (0.03) |

| Given incorrect at first test, familiar with correct answer, and response confidence | *P*(correct at retest) | Given incorrect at first test, unfamiliar with correct answer, and response confidence | *P*(correct at retest) |
|---|---|---|---|
| Omit | 0.63 (0.04) | Omit | 0.16[a] (0.05) |
| Low | 0.79 (0.03) | Low | 0.31[a] (0.08) |
| Med | 0.78 (0.05) | Med | 0.25[a] (0.07) |
| High | 0.90 (0.03) | High | 0.41[a] (0.12) |

[a] Mean of the 12 subjects with data in all cells.

a significant relationship between first test confidence and the retest accuracy, regardless of whether omit trials were included as the lowest confidence level, $F(3,57)=28.8$, $P<0.001$, or not, $F(2,38)=7.3$, $P<0.005$. Omit trials were *less* likely to be corrected at retest than were all other error types. In contrast, high-confidence errors were *more* likely to be corrected at retest than all other error types, replicating the hypercorrection effect [6].

Although subjects were on average familiar with more than half of the correct answers overall, M=0.60, S.D.=0.12, familiarity also differed as a function of first test response confidence, regardless of whether the omit trials were included in this analysis, $F(3,57)=31.9$, $P<0.001$, or not, $F(2,38)=7.5$, $P<0.005$. Specifically, for items incorrect at first test, higher confidence ratings were more likely to be followed by answer feedback subjects rated as familiar, than were lower confidence ratings. Familiarity with the correct answer also made a significant contribution to error correction in general (familiar corrected, M=0.72, S.D.=0.16; unfamiliar corrected, M=0.22, S.D.=0.16; $t(19)=29.6$, $P<0.001$). Because high-confidence errors were more likely to be followed by familiar feedback than were low-confidence errors, it is logical to infer that familiarity contributed to the hypercorrection of high-confidence errors.

Nonetheless, familiarity could not fully account for the hypercorrection effect. A significant relationship between response confidence and retest accuracy was found even when we limited this analysis to familiar correct answers (see Table 1), regardless of whether omit trials were included, $F(3,57)=13.5$, $P<0.001$, or not, $F(2,38)=4.7$, $P<0.05$. Specifically, in accordance with the hypercorrection effect, post-hoc comparisons demonstrated that these high-confidence errors were more likely to be corrected than all other error types. A similar analysis including only unfamiliar correct answers was also conducted, although only 12 participants had data in all four cells for unfamiliar trials (because high-confidence errors were rare in general, and rarely were followed by unfamiliar feedback). An ANOVA performed on data from these 12 participants found a marginal effect of confidence among unfamiliar items when omit trials were included, $F(3,33)=2.8$, $P=0.05$, stemming largely from a difference between omits and low- and high-confidence categories. No overall effect of confidence was found when omit trials were excluded, $F(2,22)=1.3$, $P=0.29$.

### 2.2.2. ERP results

*2.2.2.1. Relationship between first test accuracy and confidence.* Grand mean waveforms at the time of feedback, averaged as a function of confidence, are shown for correct responses in Fig. 1, and for incorrect responses in Fig. 2.

A 3 (confidence: low, medium, high)×2 (accuracy: correct, incorrect) ANOVA revealed main effects of both accuracy, $F(1,19)=15.8$, $P=0.001$, and confidence, $F(2,38)=4.0$, $P<0.05$, $\varepsilon=1.0$. The ERN was significantly more negative for incorrect feedback than for correct feedback and for high-confidence responses than for low- and medium-confidence responses regardless of accuracy. Although the interaction between accuracy and confidence was not significant, $F(2,38)=1.19$, $P=0.32$, the confidence effect in this epoch appeared to be driven more by variation in the response to correct feedback, which evidenced little or no ERN. A single-factor ANOVA limited to correct feedback indicated a main effect of confidence, $F(2,38)=4.24$, $P<0.05$, which resulted from an attenuation in amplitude of the ERN associated with high-confidence correct responses that was present from 200 to 500 ms. In contrast, a single-factor ANOVA limited to incorrect feedback found no main effect of confidence, regardless of whether omit responses were included as the lowest confidence level or not (all $F<1$).

Nonetheless, given that the effect of confidence observed for correct responses appeared to reflect the strong

influence of the subsequent positive component, the putative P3a, we considered the possibility that the null effect of confidence of the ERN to incorrect feedback may also have resulted from the influence of the P3a. In an attempt to minimize this influence, we calculated difference waves between incorrect and correct conditions that had similar stimulus probability and had elicited P3as of a similar size. Specifically, we subtracted low-confidence corrects from high-confidence errors (conditions featuring high metamemory mismatch and a large P3a) and subtracted high-confidence corrects from low-confidence errors (conditions featuring low metamemory mismatch and smaller P3a). These difference waves are shown in Fig. 3. The amplitude of these two difference waves was measured in the same 50 ms epoch in which we had assessed the 'raw' ERNs. Although the ERN difference waves appear numerically different, due to individual variability these values did not differ significantly, $F(1,19)=1.6$, $P=0.23$. The small negativity apparent between the P2 and P3a in the low-confidence correct responses also may have slightly reduced the amplitude of the difference wave for unexpected errors (high-confidence errors − low-confidence corrects).

Analysis of the P3a revealed a significant three-way interaction between site, accuracy, and confidence, $F(2,38)=4.9$, $P<0.05$. To explore this interaction, frontal and parietal sites were analyzed separately. At FCz, the P3a did not demonstrate significant main effects of either subjects' response accuracy or confidence (all $F<1$). However, we found a significant accuracy × confidence interaction, $F(2,38)=11.9$, $P<0.001$, $\varepsilon=0.95$. To further assess this interaction, a single-factor ANOVA limited to *incorrect* feedback found that this waveform was largest following high-confidence responses regardless of whether omits were included in the analysis, $F(3,57)=14.3$, $P<0.001$, $\varepsilon=0.76$, or not, $F(2,38)=6.8$, $P<0.005$, $\varepsilon=0.94$. In contrast, a similar analysis limited to *correct* feedback found that it was larger following low- and medium-confidence than high-confidence responses, $F(2,38)=7.5$, $P<0.005$, $\varepsilon=0.84$. At Pz, no significant effects of accuracy, $F(1,19)=2.8$, $P=0.15$, confidence, $F(2,38)=2.6$, $P=0.08$, or accuracy × confidence interaction, $F=1.0$, were found.

The broader P3b evidenced a significant site × accuracy interaction, $F(1,19)=13.7$, $P<0.005$, $\varepsilon=0.97$, which we explored by analyzing the effect of accuracy at each site separately. Incorrect feedback elicited greater positivity than correct feedback at Pz, $F(1,19)=17.0$, $P=0.001$, but there was no significant accuracy effect at FCz, $F(1,19)=2.0$, $P=0.18$.

The prominent negativity that was found over inferior temporal sites was also analyzed for the effects of accuracy and confidence. A 3 (electrode) × 2 (hemisphere) × 2 (accuracy) × 3 (confidence) ANOVA found a significant accuracy by confidence interaction, $F(2,38)=5.2$, $P=0.01$, $\varepsilon=0.84$. This interaction was driven by low- and medium-confidence corrects eliciting larger negative potentials than high-confidence corrects, whereas for incorrect responses there were no differences as a function of confidence. A single-factor ANOVA on incorrect responses that included omit trials found an effect of confidence, $F(3,57)=6.9$, $P<0.001$, $\varepsilon=0.79$, such that feedback to omit trials elicited less negativity than did feedback to other types of errors. The overall analysis also found main effects of hemisphere, $F(1,19)=5.1$, $P<0.05$, such that the negativity was larger on the left side, and site, $F(2,38)=4.7$, $P<0.05$, $\varepsilon=0.88$, such that the negativity was larger at TP9/TP10 and CB1/CB2 than at FT9/FT10.

*2.2.2.2. Relationship between familiarity with the correct answer and error correction at retest.* To assess the effect of domain familiarity on subsequent error correction, we analyzed the relationship between familiarity with the correct answer (at first test) and retest accuracy, which is illustrated by the grand means in Fig. 4. The ERN (measured from the raw waveform) did not exhibit signifi-

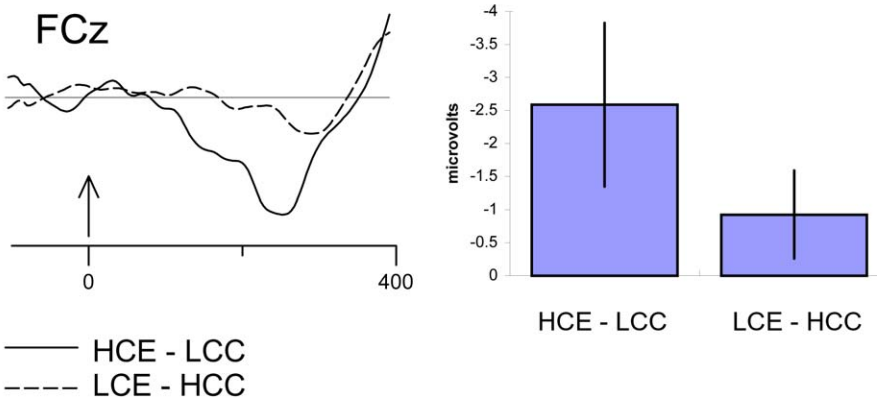


Fig. 3. Left: the ERN difference waveforms for high-confidence errors minus low-confidence corrects (HCE − LCC) and low-confidence errors minus high-confidence corrects (LCE − HCC). Right: bar graphs of the mean amplitude (during the 50 ms centered on the ERN peak-pick latency) for each difference wave with inter-subject S.E.M. bars.

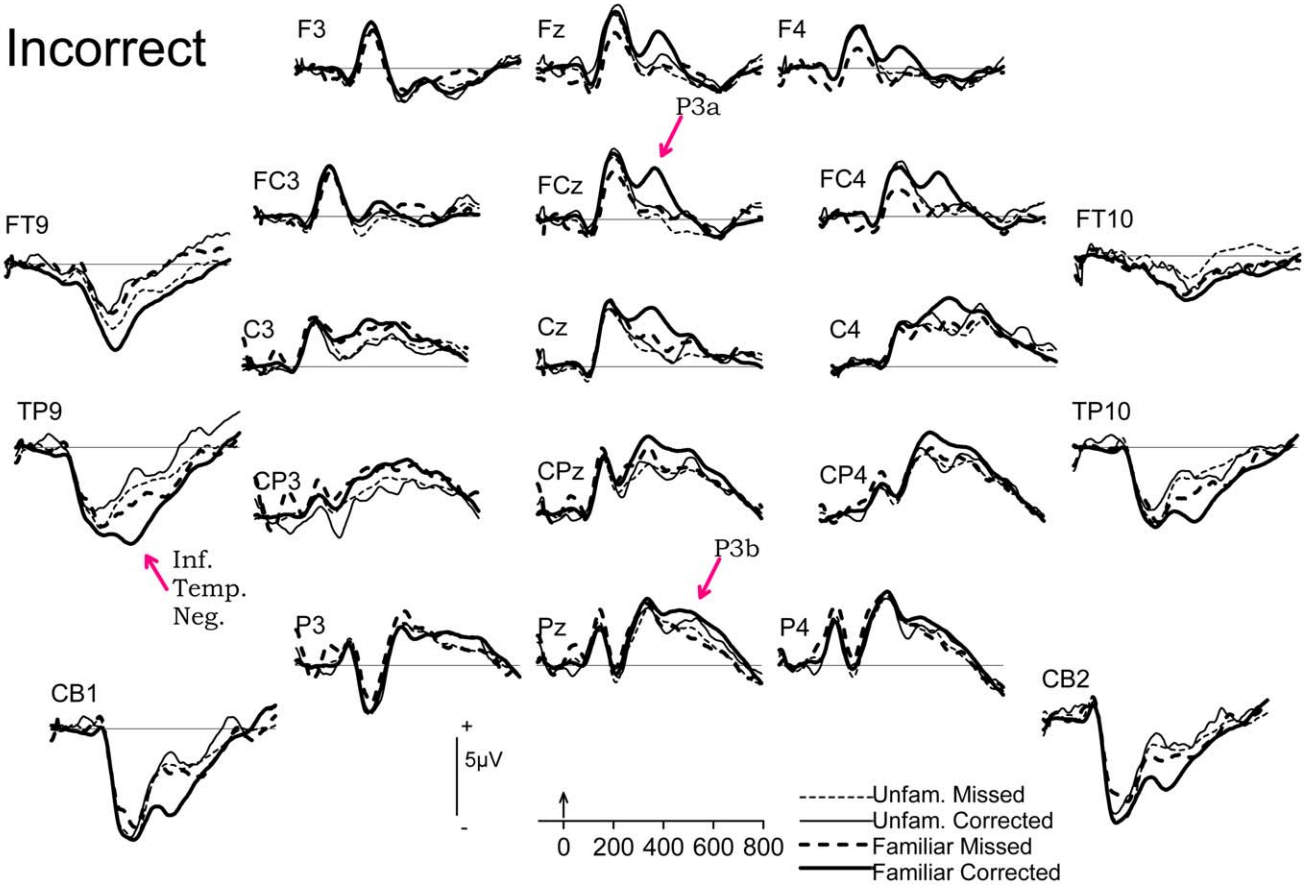


Fig. 4. Grand-mean waveforms at selected electrodes elicited by negative feedback (feedback following responses *incorrect* at first test) in Experiment 1, sorted by subjects' rating of familiarity with the correct response and subsequent retest accuracy.

cant effects of familiarity ($F<1$), error correction, $F(1,19)=1.9$, $P=0.18$, or their interaction ($F<1$).

In contrast, the P3a was largest when the subject was familiar with the correct answer, $F(1,19)=6.9$, $P<0.05$, and reliably predicted error correction on the subsequent retest, $F(1,19)=5.4$, $P<0.05$. There was a marginal main effect of site, $F(1,19)=4.1$, $P=0.06$, suggesting that the waveform was larger at the posterior (Pz) site. No interactions between site, familiarity, and error correction were found (all $F<1.7$). Although it may seem unusual that the P3a was not significantly larger at FCz, the omit and low-confidence trials, which evidenced a smaller P3a at frontal sites than high- or medium-confidence trials, made a substantial contribution to each of the error correction conditions, particularly those associated with unfamiliar answers. Analysis of the P3b also found a marginal effect of site, $F(1,19)=3.5$, $P=0.08$, also in the direction of greater amplitude at the posterior (Pz) site. Memory effects in this epoch were marginally significant [subsequent memory, $F(1,19)=3.9$, $P=0.06$; site×subsequent memory interaction, $F(1,19)=2.4$, $P=0.13$]. No other effects were significant (all $F<1$).

A strong and long-lasting relationship between familiarity and subsequent memory was found at the inferior temporal sites, however. The negative-going activity in this region was larger for familiar feedback, $F(1,19)=10.9$, $P<0.005$, and as indicated by a significant three-way interaction between hemisphere, subsequent memory and familiarity, $F(1,19)=10.0$, $P=0.005$, larger still over the left hemisphere for items that were corrected at retest. The negativity was also significantly modulated by the four-way interaction of hemisphere, electrode, familiarity, and subsequent memory, $F(2,38)=6.3$, $P=0.01$, $\varepsilon=0.72$. Post-hoc comparisons revealed that the three-way hemisphere, familiarity, and memory interaction was stronger as electrode site moved more anteriorly (i.e., it was stronger at FT9/FT10 than at TP9/TP10, and stronger at TP9/TP10 than at CB1/CB2). Reassuringly, we also replicated these results when the analysis was restricted to omit trials, $F(1,18)=7.0$, $P<0.05$ (three-way interaction, one subject who lacked sufficient omit trials, one of the four critical conditions, was excluded from analysis). Thus indicating that the relationship of the inferior temporal negativity to familiarity and subsequent memory correction persists even when the contribution of metamemory mismatch is minimized. In contrast, the P3a did not demonstrate a

reliable effect of familiarity, $F(1,18)=4.1$, $P=0.06$, or memory, $F(1,18)=2.8$, $P=0.11$, when analysis was restricted to omit responses.

### 2.3. Discussion

The behavioral findings were consistent with those of Butterfield and Metcalfe [6] in that higher confidence errors were more likely to be corrected at retest than were lower confidence errors (i.e., hypercorrection effect). The hypothesized role of familiarity in the hypercorrection effect also was supported. Questions eliciting higher confidence errors were more likely to have familiar correct answers than questions eliciting lower confidence errors, and these familiar correct answers were more likely to be generated at retest than were unfamiliar answers. However, the finding that hypercorrection could be observed even when familiarity was held constant across levels of confidence leaves open the possibility that other factor(s) are underlying this effect. We propose that one such factor is 'metamemory mismatch'—the mismatch between expected and actual performance outcome that is associated with feedback to those erroneous responses that are endorsed as correct with high confidence—and the modulation of arousal and conflict resolution processes associated with detection of that mismatch.

A positive deflection that peaked approximately 350 ms post-stimulus over fronto-central sites behaved as an electrophysiological correlate of metamemory mismatch, as it was largest under conditions in which this mismatch was greatest (i.e., high-confidence errors and low-confidence corrects) and was smallest when mismatch was smallest (i.e., low-confidence errors and high-confidence corrects). Although a positive waveform was also observed over parietal sites at this time, this positivity did not appear to be affected by metamemory mismatch. Notably, fronto-central positivity, which we descriptively term the P3a because of its earlier latency and more frontal distribution than the P3b, differs from the $P_E$ in that it is not elicited exclusively by errors in task performance, but also by errors in the subjective assessment of performance accuracy. It is interesting that the P3b in this study did not appear to be sensitive to metamemory mismatch, even though mismatches were less probable events than metamemory matches. The P3b was larger for errors overall, however, perhaps reflecting a greater need for updating of stimulus information in working memory following negative feedback than positive feedback, regardless of metamemorial expectation.

The P3a was not specific to task errors, but was related to error correction in that it was larger for items later generated at retest than items later missed, regardless of familiarity with the correct answer. Thus, to the extent that the P3a indexes the orienting of attention, this orienting appears to facilitate error correction. Furthermore, we can infer that this orienting response was particularly strong for high-confidence errors, thereby exerting a significant influence on their hypercorrection. The relationship of the P3a to error detection and correction contrasts with that of the immediately preceding negative deflection. This negativity was sensitive to task error, as we would expect if it were homologous to the ERN. However, in contrast to predictions of the Holroyd and Coles model [39], it did not appear to be modulated by the confidence with which the erroneous response had been endorsed. Failure to obtain a significant relationship between the ERN and confidence may have been due to the fact that the overlapping P3a evidenced a confidence effect in the opposite (positive) direction, yet our analysis of the difference waves designed to reduce the influence of the P3a did not reveal a significant effect either. The ERN also was not predictive of subsequent error correction. Taken together, these results suggest that the ERN indexed a fast error detection process, while the positivity that immediately followed it indexed the actual mobilization of processes that could serve to correct that error.

The influence of familiarity on the hypercorrection of high-confidence errors also was supported. The amplitude of the broad negative-going waveform over inferior temporal sites was larger for familiar than unfamiliar items, and particularly enhanced for those familiar items that could also be successfully generated at retest. This suggests that the processes indexed by this waveform are involved in the activation of pre-existing information in semantic memory and/or processes involved in strengthening or creating an association between the question and the familiar answer. The finding that this waveform was larger overall to feedback following correct as compared to incorrect responses supports this interpretation. All the correct responses produced by the subject would be familiar already, whereas feedback to incorrect responses includes both familiar and unfamiliar correct answers. In addition, in the case of correct responses, the correct answer was recently activated in semantic memory, which would further increase its familiarity. Interestingly, for correct responses, but not incorrect responses, the inferior temporal negativity was modulated by confidence, in that it was larger for low-confidence trials than for high-confidence trials. This pattern could also be viewed as consistent with the idea that this waveform indexes a process that integrates associations between familiar elements, either for the purpose of establishing a new question–answer association, as was necessary in the case of all error types, or for strengthening an extant but weak association, as was the case only for low-confidence corrects.

The finding that the inferior temporal negativity was significantly smaller to feedback following omit responses than non-omit errors provides a potential caveat to this interpretation, however. The reduced amplitude of the inferior negativity to omit trials may be due, in part, to the fact these trials were the least likely error type to be associated with a familiar correct answer (51% familiar),

as well as error type least likely to be corrected (41% corrected). However, it is unlikely that this explanation is sufficient, as the difference in familiarity and error correction between omits and low-confidence errors (68% familiar, 63% corrected) was comparable to the difference between low-confidence errors and high-confidence errors (79% familiar, 82% corrected), and these latter error types did not differ from each other at this waveform. Thus, there may be another contributor to this difference. This factor may be attention.

Again, to the extent that the P3a indexes an orienting response, its attenuation following omit responses suggests participants did not orient a great deal of attention to feedback in this condition, or at least oriented less than when they had attempted an answer. Similarly, attenuation of this inferior temporal negativity for omit responses may reflect reduced attention to semantic processing of the correct answer (see also Ref. [58]), perhaps because participants were less invested in learning the correct answer to questions where they had opted to make an omit response. Indeed, the finding that participants succeeded in correcting fewer items in the 'familiar' omit condition than for other levels of confidence suggests that even when the correct answer was recognized as familiar, the association between that familiar answer and the question was less likely to be successfully strengthened. This failure may be the result of reduced attention to this process.

In addition to the differential effects of confidence on the P3a and inferior temporal negativity, there are other differences between these waveforms that support the view that the inferior temporal negativity indexes processes associated with familiarity and subsequent memory that are independent of any potential overlap with the inverse of the dipole responsible for the P3a. First, although the inferiority temporal negativity and P3a were modulated by subsequent memory, the subsequent memory effect found over inferior temporal regions appeared to extend longer into the epoch (600 ms) than the P3a (475 ms). More compelling perhaps was the functional discussion found when we limited our analysis to omit responses. In this case, the inferior temporal negativity remained significantly larger for corrected than for uncorrected items, but these effects were no longer robust for the P3a. It is also unlikely that the inferior negativity simply represents the inverse of the posterior P3b, despite the similarity of their time courses and the sensitivity of similar P3 waveforms to memory in other studies (e.g., Ref. [20]). In this study, the P3b only demonstrated at best a weak relationship to either subsequent memory or familiarity (see also Ref. [58]).

Nonetheless, because presentation of the accuracy feedback (color) and the correct answer feedback (word) were simultaneous, there are methodological and theoretical difficulties in fully isolating the cognitive constructs associated with the inferior temporal negativity and the two overlapping positive waveforms. In addition to the issues described above, it is difficult to compare the extent to which the P3a was sensitive to metamemory mismatch versus familiarity because familiarity and confidence in an error were positively correlated. Moreover, the process indexed by the ERN also is somewhat ambiguous, given that the ERN could be indexing either error detection and/or the conflict between the participant-generated incorrect response and the presented correct response. We attempted to address these issues in Experiment 2 by separating the feedback into an accuracy component (color and tone) and a correct answer component (correct answer presented in white). In this way, we could potentially separate processes associated with error detection and metamemory mismatch (i.e., neural response to accuracy feedback), from processes associated with familiarity, semantic association and response conflict (i.e., neural response to correct answer).

Experiment 2 also attempted to address the issue of encoding versus consolidation. Given that the retest in Experiment 1 took place after a relatively short delay, we cannot say whether the associations between these corrected answers and their question context have been (or will be) permanently integrated into long-term declarative memory and exhibit characteristics of a stable semantic or episodic memory. Furthermore, although this study replicates the hypercorrection of high-confidence errors at an immediate retest, it is not clear whether this effect would be reduced or enhanced by a longer study-test delay. Indeed, it is possible that the difference between correction of high- and low-confidence errors would be even greater after a delay than at immediate test. Specifically, negative feedback to high-confidence errors may elicit greater emotional arousal than other errors because of the embarrassment associated with metamemory mismatch, and, therefore, may activate the amygdala, which may facilitate consolidation of episodic memory for the corrective information and the likelihood that this information will be recalled even after an extended delay (e.g., Refs. [8,9,38]). Thus, in Experiment 2, we included a week-delayed retest to see if the hypercorrection of high-confidence errors, and its relation with the ERP components described above, persists over time.

## 3. Experiment 2

### 3.1. Method

#### 3.1.1. Participants

Twenty-three participants (12 females, mean age 20 years) were tested who were screened for first language, sensory, neurological and/or psychological disorder and trivia knowledge in the same way as in Experiment 1. Data from three participants were lost or excluded due to computer problems or excessive noise. All participants gave informed consent and were compensated at a rate of $10/h for their participation.

*3.1.2. Materials*

The stimuli were 680 trivia questions from a variety of knowledge domains. As in Experiment 1, questions had answers that were single words three to eight letters in length. These questions included the questions from Experiment 1 as well as 460 taken from various Internet trivia sites and trivia board games. Each participant, however, was presented with only 350 of these questions.

*3.1.3. Procedure*

*3.1.3.1. Trivia question task.* The experiment consisted of three phases: a test, a surprise retest, and a week-delayed retest. EEG was recorded during the first test phase only. Trivia questions were presented in the center of the computer screen and the participant was given an unlimited amount of time to type his/her response on the computer keyboard. To increase the likelihood that a given participant would have a large number of trials in our conditions of interest, trivia questions selection was accomplished on-line with an adaptive algorithm by which normatively more difficult questions were more likely to be presented if the participant's current performance was above 40% correct, and normatively less difficult questions were more likely to be presented if the participant's current performance was below 40% correct. The details of this algorithm can be found in Appendix A. Participants were instructed to provide responses and rate their confidence in the accuracy of their responses the same way as in Experiment 1.

Immediately following the confidence rating or 'xxx' (omit) response, a central fixation point appeared for 1.5 s. The first feedback was then provided for 1 s in the form of a cross and a tone. The cross was presented in green with a high-pitched tone if the participant's response was correct, and in red with a low-pitched tone if it was incorrect. A second fixation cue was then presented for 1.5 s, followed by feedback in the form of the correct answer presented for 2.5 s. Participants were instructed to avoid blinking or moving during both feedback periods. The first feedback to 'xxx' responses was always a red cross with a low tone, as was the feedback to borderline responses, however these latter trials were not included in any analyses. Following the second feedback (i.e., the correct answer) after incorrect or borderline trials, participants indicated their familiarity with the correct answer. Participants could respond that it was unfamiliar or familiar, as in Experiment 1, or whether they 'knew it,' meaning that at the time they received the negative feedback signal, they had realized immediately what the correct answer was and presentation of this answer merely confirmed their prediction. In contrast, a response of 'familiar' simply meant that the correct answer was familiar, but that they had not predicted it at the time of the accuracy feedback. Participants were given short breaks after the 87th, 175th, and 262nd trial.

There was a delay of approximately 8 min after the first test, during which time the recording electrodes were removed. Participants then began the first retest phase, which consisted of a surprise retest for a random half of the questions answered incorrectly in the first test phase. A week later, participants returned and completed the second retest phase, which consisted of the other half of the questions answered incorrectly in the first test phase. To ensure that the number and type of items retested immediately and at the week-delayed retest were roughly equal, questions that were incorrect at first test were sorted first by familiarity rating ('knew it,' familiar, and then unfamiliar) and then by confidence. Items in each successive pair of items in this list was randomly assigned to the immediate or week-delayed retest. If there were an odd number of incorrect trials, the remaining item was randomly designated as an immediate or week-delayed retest item. In both the immediate and the week-delayed retest, the sequence of trial events in the retest phase was the same as during the first test, with the exceptions that neither response confidence nor familiarity were assessed, and feedback was in the form of the correct answer in green (if correct) or red (if incorrect or borderline).

*3.1.3.2. Electrophysiological recording.* The ERP recording method was the same as in Experiment 1.

*3.1.3.3. ERP analysis.* ERPs were time-locked to the first feedback (color and tone) and the second feedback (correct answer). Out of a total of 7000 epochs (350×20 participants), 28 (<1%) were rejected because of a behavioral response of borderline accuracy either at first test or either retest. Of the remaining 6972 trials, 515 (<8%) were rejected due to excessive noise.

As in Experiment 1, we collapsed the seven-point confidence scale into low (1–3), medium (4), and high (5–7) confidence categories. Due to the greater number of first-test trials, all levels of confidence now had an average of at least 10 trials per participant. However, we were forced to limit our error correction analysis to familiar items because relatively few items were rated as unfamiliar (M=26%), and those that were rated unfamiliar were rarely recalled at the week-delayed retest (M=9.8%). Indeed, five participants failed to correctly generate any correct unfamiliar trials at this delay. All averages excluded items where the participant claimed to 'know' the correct answer as soon as the accuracy feedback was presented. These items represented a categorically different type of response than either familiar or unfamiliar items, were relatively rare (M=11.2%), and could have been contaminated by hindsight bias.

All effects were analyzed with ANOVAs using the mean amplitude as the dependent variable. Mean amplitude windows were centered on the peak latency. At FCz, the mean latency of the ERN to the accuracy feedback was 262 ms. When ERN latency was subjected to a 2 (experiment)×2 (accuracy)×3 (confidence) mixed

ANOVA, it was found to be marginally earlier than the ERN observed in Experiment 1, which had peaked at 276 ms, $F(1,38)=3.4$, $P=0.07$). Therefore, in the present experiment, we analyzed the ERN at FCz using the mean amplitude from 236 to 284 ms. Similarly, the P3a to accuracy feedback also peaked earlier at FCz (320 ms) than in Experiment 1, $F(1,38)=17.6$, $P<0.001$. Therefore, we analyzed this waveform using the mean amplitude between 296 and 344 ms. For analysis of the P3b, we used the mean amplitude within broader windows based on when this waveform appeared maximal in the grand means; 400–600 ms for the accuracy feedback, and 500–700 ms for the answer feedback. As in Experiment 1, P3a and P3b waveforms were analyzed at both FCz and Pz. The inferior temporal negativity was also analyzed in different portions of the epoch for accuracy and answer feedback, based on when this effect was maximal in the grand means; accuracy feedback: 296–244 ms (same as the P3a), answer feedback: 300–600 ms. It was analyzed at the same electrode pairs as in Experiment 1.

Greenhouse–Geiser corrections [41] were applied where appropriate. Epsilon values will be reported alongside uncorrected degrees of freedom. Significant effects were investigated with Tukey's HSD post-hoc comparisons. The alpha level for all analyses was 0.05.

### 3.2. Results

#### 3.2.1. Behavioral results

Our adaptive algorithm for selecting question difficulty proved successful in maintaining participants' first test accuracy at around 40%, M = 0.36, S.D. = 0.07. Moreover, as shown in the first column of Table 2, participants' confidence ratings continued to be reliable indicators of first-test accuracy, $F(2,38)=580.2$, $P<0.001$. Table 2 also illustrates the probabilities of responses at each confidence level as a function of accuracy at first test and each retest, being associated with familiar or 'known' correct answers, and being correct at retest as a function of answer familiarity.

To assess the relationship between first-test confidence and retest accuracy, we conducted an ANOVA analyzing accuracy at immediate and delayed retest as a function of confidence at first test. A main effect of confidence was found regardless of whether omits were included, $F(1,19)=86.0$, $P<0.001$, or not, $F(1,19)=15.8$, $P<0.001$. Post-hoc comparisons indicated this main effect was due to the hypercorrection of high-confidence errors relative to all other response categories. A main effect of retest delay also was found whether omits were included, $F(1,19)=138.9$, $P<0.001$, or not, $F(1,19)=102.8$, $P<0.001$, such that more errors were corrected at immediate retest than week-delayed retest. However, hypercorrection did not differ as a function of retest delay, as there was no interaction between confidence and retest delay, $F<1$.

For items incorrect at first test, there was a significant relationship between response confidence and the likelihood that the correct answer was familiar when omits were included in the analysis, $F(3,57)=24.7$, $P<0.001$, but not when they were not included, $F(2,38)=1.7$, $P=0.20$. The lack of an effect when omits were not included in the analysis appeared to be due to an increasing number of items being rated as 'known' at higher confidence levels. Indeed, when the analysis was computed on the probability of the feedback being familiar or known a significant effect was obtained, $F(2,38)=21.7$, $P<0.001$. Pairwise comparisons found the likelihood of feedback familiarity was greater for high- and medium-confidence errors than for low-confidence errors.

Familiarity also made a significant contribution to error correction. Overall, participants were more likely to correct their answer when that answer was already familiar (not

Table 2
Conditional probabilities (with S.E.M.) of responses in Experiment 2

| *P*(response confidence) | | Given resp. conf. | *P*(correct at first test) | Given wrong at first test and resp. conf. | *P*(correct at immed. test) | *P*(correct at delayed test) | *P*(correct answer) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Familiar | Known |
| Omit | 0.21 (0.03) | Omit | – (–) | Omit | 0.56 (0.03) | 0.28 (0.02) | 0.54 (0.03) | 0.04 (0.01) |
| Low | 0.31 (0.02) | Low | 0.17 (0.02) | Low | 0.76 (0.02) | 0.49 (0.03) | 0.72 (0.03) | 0.08 (0.01) |
| Med | 0.15 (0.02) | Med | 0.36 (0.03) | Med | 0.88 (0.03) | 0.58 (0.04) | 0.77 (0.03) | 0.14 (0.02) |
| High | 0.33 (0.02) | High | 0.78 (0.02) | High | 0.87 (0.02) | 0.61 (0.04) | 0.74 (0.03) | 0.16 (0.03) |

| Given incorrect at first test, familiar with correct answer, and response confidence | *P*(correct at immed. test) | *P*(correct at delayed test) | Given incorrect at first test, unfamiliar with correct answer, and response confidence | *P*(correct at immed. test) | *P*(correct at delayed test) |
|---|---|---|---|---|---|
| Omit | 0.78 (0.03) | 0.40 (0.04) | Omit | 0.25[a] (0.05) | 0.07[b] (0.02) |
| Low | 0.84 (0.02) | 0.55 (0.03) | Low | 0.37[a] (0.10) | 0.15[b] (0.06) |
| Med | 0.93 (0.02) | 0.60 (0.04) | Med | 0.58[a] (0.16) | 0.20[b] (0.11) |
| High | 0.90 (0.02) | 0.59 (0.06) | High | 0.37[a] (0.15) | 0.15[b] (0.11) |

[a] Mean of the seven subjects with data in all cells.
[b] Mean of the nine subjects with data in all cells.

including known) than when it was unfamiliar at both immediate retest [M=0.83 (S.D.=0.08) vs. M=0.31 (S.D.=0.17), $t(19)=17.5$, $P<0.001$] and week-delayed retest [M=0.53 (S.D.=0.13) vs. M=0.10 (S.D.=0.10), $t(19)=15.4$, $P<0.001$]. A retest delay×confidence ANOVA limited to familiar items (see Table 2) found a main effect of confidence regardless of whether omits were included in the analysis, $F(3,57)=18.4$, $P<0.001$, or not, $F(2,38)=4.6$, $P<0.05$. This analysis also found a main effect of retest delay, such that more errors were corrected at immediate retest than at delayed retest, with omit trials $F(1,19)=141.4$, $P<0.001$, or without omit trials, $F(1,19)=83.6$, $P<0.001$. There was no interaction of retest delay with confidence ($F<1$). We chose not to conduct a similar analysis of unfamiliar items because of low power due to the highly reduced number of subjects providing any trials in these conditions, as well as the small numbers of trials that these few subjects did provide. Nonetheless, we show the mean data for this group in Table 2 for descriptive purposes.

### 3.2.2. ERP results, accuracy feedback

*3.2.2.1. Relationship between first-test accuracy and confidence*. Fig. 5 illustrates the grand mean waveforms averaged as a function of confidence at electrodes selected to highlight the effects of interest. A 3 (confidence: low, medium, high)×2 (first-test accuracy: correct, incorrect) ANOVA indicated that the ERN was significantly more negative following negative feedback (incorrect responses) than following positive feedback (correct responses), $F(1,19)=29.4$, $P<0.001$. It was also larger overall for high-confidence responses than for low- and medium-confidence responses, $F(2,38)=3.5$, $P<0.05$, $\varepsilon=0.85$, however this finding is qualified by a significant interaction between accuracy and confidence, $F=6.8$, $P<0.005$, $\varepsilon=0.85$. Single-factor ANOVAs confirmed that this confidence effect occurred only for correct responses, which evidenced little or no ERN, $F(2,38)=7.9$, $P=0.005$, $\varepsilon=0.69$. As in Experiment 1, this confidence effect appeared to result from an overall reduction in positivity for high-confidence correct items that started as early as 200 ms and extending through to 500 ms. Also replicating the results of Experiment 1, we found no effect of confidence on the ERN to incorrect responses, $F(2,38)=1.4$, $P=0.27$, $\varepsilon=0.69$. Yet, when we analyzed the ERN using difference waves designed to minimize the effects of the leading edge of the P3a (see Experiment 1), significant results were obtained. Specifically, the difference wave for high-confidence errors minus low-confidence corrects was significantly larger than the difference wave for low-confidence errors minus high-confidence corrects, $F(1,19)=5.4$, $P<0.05$ (see Fig. 6).

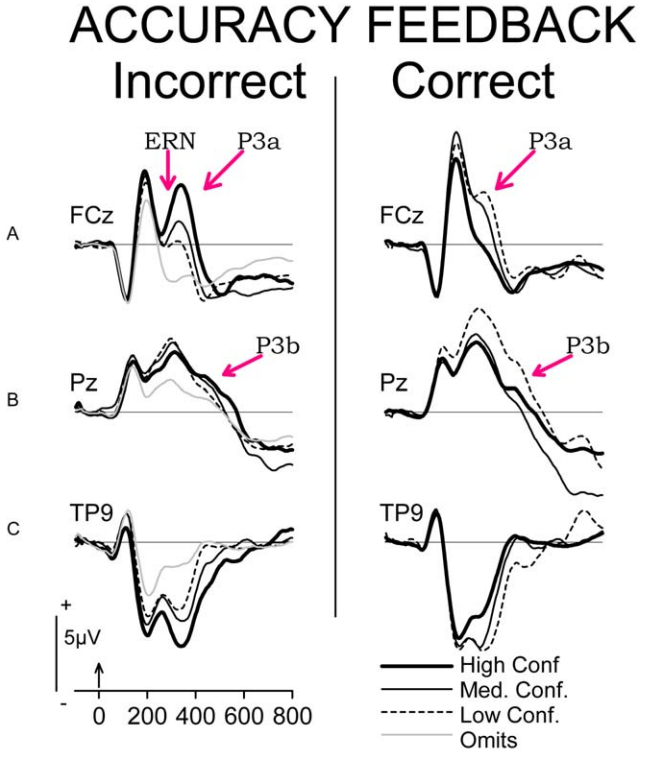


Fig. 5. Grand-mean waveforms for the accuracy feedback to responses in Experiment 2, sorted by first test accuracy and confidence. Data is shown at sites (A) FCz, (B) Pz, and (C) FT9.

The P3a evidenced a main effect of site, $F(1,19)=7.7$, $P<0.05$, interaction between accuracy and confidence, $F(2,38)=9.5$, $P<0.001$, $\varepsilon=0.83$, and interaction between site, accuracy, and confidence, $F(2,38)=7.9$, $P=0.001$, $\varepsilon=0.83$. The site×accuracy interaction did not reach significance, $F(1,19)=1.9$, $P=0.18$. Therefore, we pursued the three-way interaction by analyzing the effects of accuracy and confidence at FCz and Pz separately. At FCz, the P3a was not modulated by either accuracy or confidence ($F<1$), but was strongly modulated by their interaction, $F(2,38)=19.0$, $P<0.001$, $\varepsilon=0.83$. Single factor ANOVAs confirmed that negative accuracy feedback elicited a larger P3a following higher-confidence responses than following lower-confidence responses, regardless of whether omits were included, $F(3,57)=19.9$, $P<0.001$, $\varepsilon=0.53$, or not, $F(2,38)=12.0$, $P=0.001$, $\varepsilon=0.69$, whereas for positive accuracy feedback, lower-confidence responses elicited a larger P3a than did higher-confidence responses, $F(2,38)=13.5$, $P<0.001$, $\varepsilon=0.78$. Activity at Pz was reliably modulated by confidence overall, $F(2,38)=3.5$, $P<0.05$, $\varepsilon=0.97$, although post-hoc comparisons (which corrected for multiple comparisons) did not find any significant pairwise differences. There were no significant main effects of accuracy or interaction of accuracy×confidence at this site ($F<1.4$).

A similar analysis of the broader P3b found that this waveform was larger at Pz than at FCz, $F(1,19)=6.1$, $P<0.05$. There was a weak effect of confidence, $F(2,38)=2.1$, $P=0.13$, and a marginal interaction of site×accuracy,

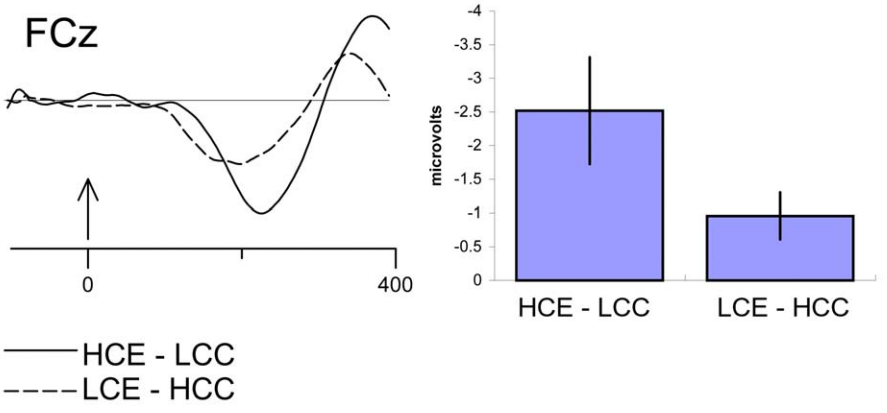


Fig. 6. Left: the ERN difference waveforms for high-confidence errors minus low-confidence corrects (HCE−LCC) and low-confidence errors minus high-confidence corrects (LCE−HCC). Right: bar graphs of the mean amplitude (during the 50 ms centered on the ERN peak-pick latency) for each difference wave with inter-subject S.E.M. bars.

$F(1,19)=3.6$, $P=0.07$. No other effects were even marginally significant ($F<1.6$). In particular, we failed to observe an accuracy×confidence interaction at Pz in either this experiment or Experiment 1 despite the fact that a similar waveform has been shown to be sensitive to stimulus probability in previous studies [18]. To investigate the possibility that the sensitivity of our P3b measurement (i.e., a mean epoch) was reduced by latency jitter, we used a Woody's filter to lock averaging onto the onset of this waveform on a trial-by-trial basis [93].[2] Regardless of whether we used a fixed-size half-sinusoid template or a flexible template that maximized the cross-correlation 'fit' trial by trial, we observed no significant accuracy× confidence interaction at Pz ($F<1.5$).

A negative waveform over inferior temporal electrodes was affected in an identical manner to the P3a at FCz with regard to the accuracy by confidence interaction, $F(2,38)=30.8$, $P<0.001$, $\varepsilon=0.83$ (see Fig. 5). This interaction was more evident at posterior electrodes (TP9/TP10, Cb1/Cb2) than at frontal electrodes (FT9/FT10), $F(4,76)=11.0$, $P<0.001$, $\varepsilon=0.52$.

*3.2.2.2. Error correction at retest for familiar items*. To assess the relation of these waveforms to subsequent error correction, we averaged the ERPs elicited by the accuracy feedback according to retest accuracy and retest delay. The grand means of these averages are shown at selected electrodes in the left panel of Fig. 8.

The ERN did not exhibit reliable effects of error correction ($F<1$), retest delay, $F=3.0$, $P=0.10$, or their interaction, $F(1,19)=3.7$, $P=0.07$. The P3a demonstrated a main effect of site $F(1,19)=12.6$, $P<0.005$, but as in Experiment 1, was actually more positive overall at Pz than at FCz. In addition, this waveform demonstrated an overall effect of delay, $F(1,19)=4.9$, $P<0.05$, such that items retested immediately elicited a smaller P3a. There were no effects of error correction, site×delay, or site× error correction interactions (all $F<1.7$). A trend toward interaction of retest delay×error correction, $F(1,19)=2.9$, $P=0.10$, was qualified further by a marginal interaction between site, delay and error correction, $F(1,19)=3.6$, $P=0.07$. We tentatively went forward and addressed the three-way interaction by analyzing the effects of delay and error correction interaction at FCz and Pz electrodes separately. At Pz, there were no significant effects (all $F<1.4$). At FCz, however, the P3a was marginally related to error correction, $F(1,19)=3.0$, $P=0.10$, but robustly sensitive to retest delay, $F(1,19)=5.8$, $P<0.05$, and the interaction of error correction and delay, $F(1,19)=7.9$, $P<0.05$. This interaction was driven by the reduced amplitude of the P3a associated with errors not corrected at immediate retest, which was reliably smaller than the P3a to the other three conditions. The P3b demonstrated only a main effect of site, such that it was larger at Pz, $F(1,19)=8.3$, $P<0.01$. No other effects were significant ($F<1.6$). Single-trial analysis of the P3b using a Woody's filter identical to the one used above also found no significant effects at Pz ($F<1.7$).

Analysis of the negativity at inferior temporal sites found that it was significantly smaller for items missed at immediate retest than it was for all other types of items, and that this effect was larger over the left hemisphere. This was evidenced by a three-way hemisphere×delay×

[2]The algorithm searched for the highest cross-correlation within the 200–600 ms post-feedback epoch. The fixed template was a half-sinusoid 200 ms wide and 4 μV in amplitude. The flexible template was between 150 and 250 ms wide, in 25 ms increments, and between 3 and 9 μV high, in 1.5 μV increments. Thus, the algorithm cross-correlated all 35 possible templates across the epoch and time-locked each epoch at the beginning of the template that achieved the best correlation, where it achieved it. The average ERP was then calculated for each condition and each subject with these onset-adjusted epochs. Both methods achieved a 'humped' distribution of where they found the P3s to begin.

error correction interaction, $F(1,19)=4.9$, $P<0.05$. The negativity was also larger in general at TP9/10 and CB1/2 than it was at FT9/10, $F(2,38)=6.9$, $P<0.01$, $\varepsilon=0.70$.

### 3.2.3. ERP results, answer feedback

*3.2.3.1. Relationship between first-test accuracy and confidence*. As shown in Fig. 7, a measurable ERN was not present in the activity elicited by the correct answer feedback. The P3a was also attenuated in comparison to the waveform elicited by accuracy feedback, but could still be measured in individual subjects. The mean latency of the P3a to the answer feedback was 384 ms, which was significantly later than both the P3a to the accuracy feedback in the present experiment, $F(1,38)=69.8$, $P<0.001$, and the P3a to the combined feedback in Experiment 1, $F(1,38)=11.6$, $P<0.005$. A 2 (site)×3 (confidence)×2 (accuracy) ANOVA on the P3a demonstrated a main effect of site, such that activity at Pz was more positive, $F(1,19)=12.3$, $P<0.05$, and accuracy, $F(1,19)=7.6$, $P<0.05$, such that correct answers showed less positivity overall. There was also a marginal interaction of site×accuracy, $F(1,19)=2.9$, $P=0.10$. There were no other significant effects ($F\leq1.8$). Similarly, the P3b exhibited only an effect of site, such that it was larger at Pz, $F(1,19)=15.6$, $P=0.001$. There was a weak trend for an effect of accuracy, $F(1,19)=2.3$, $P=0.15$, but no other significant effects ($F<1$).

The correct answer elicited a larger negativity over inferior temporal sites when subjects had initially produced the incorrect response, especially over the right hemisphere, as indicated by post-hoc comparisons conducted to interpret a significant accuracy×hemisphere interaction, $F(1,19)=9.3$, $P<0.01$. There was also a main effect of electrode site, such that the negativity was larger at posterior compared to anterior sites, $F(2,38)=6.1$, $P<0.05$, $\varepsilon=0.70$.

*3.2.3.2. Error correction at retest for familiar items*. As shown in the right panel of Fig. 8, the amplitude of the P3a to the correct answer was larger at Pz than at FCz, $F(1,19)=7.5$, $P<0.05$, but was not significantly affected by subsequent memory, retest delay, or their interaction ($F<1.4$). The P3b was also larger at Pz, $F(1,19)=22.3$, $P<0.001$, but evidenced no other significant effects ($F<1$), even when a single trial analysis was conducted.[3]

In contrast, the inferior-temporal negativity exhibited a long-lasting effect of subsequent memory. From 300 to

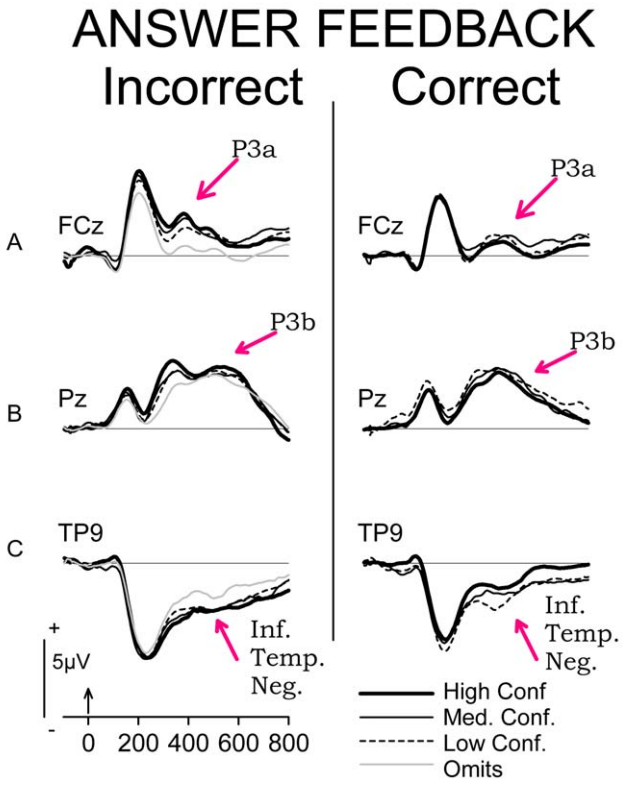


Fig. 7. Grand-mean waveforms for the answer feedback to responses in Experiment 2, sorted by first-test accuracy and confidence. Data is shown at sites (A) FCz, (B) Pz, and (C) FT9.

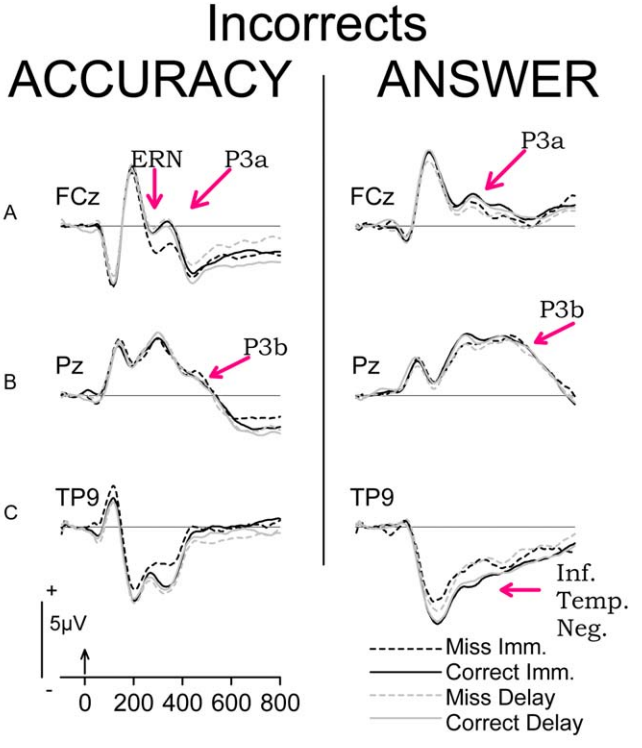


Fig. 8. Grand-mean waveforms for the accuracy feedback and the answer feedback following incorrect first-test responses in Experiment 2. Data is shown at sites (A) FCz, (B) Pz, and (C) FT9.

[3]This was identical to the other filter used, with the exception that the 'eligible' epoch was from 300 to 800 ms and the possible template widths were 200 to 300 ms wide. The distribution of acquired onsets was humped, but less so than it was for the accuracy feedback data, which suggests that the parietal P3 may be less peaked and more sustained in the answer feedback data than it is in the accuracy feedback data.

600 ms, the activity within this region was significantly more negative for familiar items corrected at retest than for those missed at retest delay, $F(1,19)=7.2$, $P<0.05$, yet did not differ by retest delay or the interaction between retest accuracy and retest delay ($F<1$). Unlike in Experiment 1, however, this effect was not reliably more pronounced over the left hemisphere ($F<1$).

### *3.3. Discussion*

Separating feedback into sequential accuracy and correct answer components successfully disambiguated the relationship between the ERN, P3 waveforms, and inferior temporal negativity, both in terms of their temporal characteristics and their relationship to error detection and correction processes. The P3a to the accuracy feedback peaked earlier than the P3a to the answer feedback, indicating that evaluation of the non-verbal accuracy feedback proceeded more quickly than evaluation of the answer feedback. More importantly, however, the P3a elicited by the composite feedback in Experiment 1 peaked between the accuracy and answer feedback in the present experiment, supporting the hypothesis that in Experiment 1 the P3a represented a 'smearing' of the processes associated with the detection of a metamemory error (triggered by the color of the word), and processes associated with evaluation of familiarity and error correction (triggered by the correct answer), as well as any interaction of mismatch and familiarity that might have occurred when participants realized that the correct answer they failed to produce was already a part of their semantic repertoire. Similarly, the sequential feedback allowed us to separate the aspect of the inferior temporal negativity that appeared to be simply the inverse of the P3a (i.e., inferior negativity following accuracy feedback) from the aspect of this component that was broader and functionally independent from the coincident positive components (i.e., inferior negativity following answer feedback).

Both the ERN and fronto-central P3a were more prominent to feedback that conveyed information about the accuracy of the response than to feedback specifying the correct answer, suggesting that these components indexed processes involved in the initial detection of the error. Unlike in Experiment 1, however, there was some evidence that the ERN in the present experiment was sensitive not only to detection of negative feedback, but also the subject's expectation regarding whether that negative feedback was going to occur. Specifically, although no effect of confidence was evident in analysis of the 'raw' waveforms, ERN difference waves that reduced the influence of the overlapping P3a found a larger ERN associated with metamemory mismatch than metamemory match. The anterior aspect of the P3a also was sensitive to mismatch between expected and actual accuracy (metamemory mismatch), but unlike the ERN, robustly demonstrated this effect regardless of whether task feedback was positive or negative. In contrast to either fronto-central waveform, the later posterior P3b did not demonstrate sensitivity to either accuracy or the interaction between accuracy and confidence.

Regarding the relationship of the P3a to subsequent memory, the pattern of results found in this experiment was somewhat complex, however it appeared that the amplitude of the anterior aspect of the P3a to accuracy feedback only differentiated whether the correct answer that followed would be recalled or forgotten at the immediate retest. These findings essentially replicate the results of Experiment 1, which also found that the P3a predicted subsequent memory at immediate retest. Thus, the failure to correct errors after a relatively short delay may be related to insufficient orienting of attention to corrective information. Yet, the amplitude of this component did not differentiate successful recall at the 1-week delay. Attention may be necessary, but not sufficient for consolidating these items in the more permanent long-term memory that was more likely to be tapped by the week-delayed retest.

Successful error correction at both the immediate and longer delay depended upon successful strengthening of the semantic relationship between the question and the correct answer, a process that was facilitated when the item was already familiar. Unlike in Experiment 1, such processes could not be engaged until the correct answer was presented, and therefore, should not have occurred concurrently with the detection of metamemory mismatch except in the rare cases when subjects claimed to have 'known' the correct answer at the accuracy feedback (these items were not included in the analysis). Indeed, although a small P3a was observed following presentation of the correct answer, this component was not predictive of subsequent memory. Rather, the broad inferior temporal negativity, which we hypothesize to index a semantic integration process, was predictive of successful recall of familiar answers at both the immediate and week-delayed retest. Although this broad waveform overlapped temporally with a measurable P3b to the answer feedback, there was no evidence that this posterior positivity was predictive to subsequent memory at either retest. Thus, these results provide further support for the independence of the processes indexed by the P3 waveforms and the inferior temporal negativity, and additionally, suggest that the process indexed by inferior temporal negativity underlies an aspect of memory formation that allows memories to endure over time.

## 4. General discussion

### *4.1. Overview*

In two experiments, we examined the sequence of cognitive and neural events underlying the detection and correction of errors in semantic knowledge. We focused in particular on how these processes are modulated when

errors in memory retrieval are accompanied by errors in metamemory (i.e., when erroneous responses are endorsed as correct with high confidence at first test). Such errors, which violate subjects' metamemorial expectations, are of particular interest because they are also associated with greater success in committing the correct answer to long-term memory—a 'hypercorrection' effect first observed by Butterfield and Metcalfe [6].

With regard to error *detection*, we observed an early fronto-central negative deflection that appeared to be specific to negative feedback, similar to the ERN.[4] The relationship of this ERN-like waveform to the violation of metamemorial expectation was less clear, however. No effects of metamemory mismatch (i.e., expectancy) were found when these ERN-like waveforms were analyzed directly as a function of response confidence. Yet, a significant modulation by expectancy was found in Experiment 2 when the overlapping effects of the large positive waveform that followed this negativity were subtracted out. Specifically, the magnitude of the ERN difference wave associated with unexpected errors (high-confidence errors − low-confidence corrects) was larger than the ERN associated with errors that were expected (low-confidence errors − high-confidence corrects). A similar pattern of results was found in Experiment 1, but the difference between conditions did not reach significance in that study.

A positive waveform subsequent to the ERN was more robustly sensitive to the degree of mismatch between expected and actual feedback, exhibiting this relationship in both experiments. The spatial and temporal distribution of positivity suggests that it is analogous to the fronto-central novelty-P3/P3a typically elicited by unexpected events in a three-stimulus (i.e., standard, target, novel) oddball task [47–49,80,81], and therefore we referred to this waveform as the P3a. Unlike the ERN, this positivity clearly indexed metamemory mismatch regardless of whether the actual outcome was positive or negative.

Although unexpected (high-confidence) errors were associated with a larger P3a and were more likely to be corrected at both immediate and delayed retest than low-confidence errors or omit responses, the P3a was only significantly related to error *correction* on the immediate retest. However, large behavioral effects of answer familiarity were found at both immediate and delayed tests, which were mirrored by a broad inferior temporal negativity that was sensitive to the familiarity of the correct answer and subsequent error correction, regardless of whether all items or only omit responses were included in the analysis, and regardless of retest delay.

### 4.2. Error detection: task error and metamemory error

The sensitivity of the earlier fronto-central negativity to error feedback provides further support for the association of the ERN to a 'generic' error-processing system that is elicited in response to detection of a mismatch between a subject's response and the goals of the task [39,61,77]. Although the presence of a small ERN following positive feedback to low-confidence correct responses (see Fig. 1) could indicate that it represents an initial response checking process that simply is enhanced when an error is detected, our results suggest that an ERN can be elicited under conditions other than those of direct response conflict. In Experiment 2, we found an ERN to accuracy feedback alone, before the correct response was presented to the subject and could have generated competition. Notably, in previous studies of the feedback-locked ERN, it was more difficult to disambiguate error detection and response conflict because the response choices were binary, such that negative feedback automatically informed the subject of the correct, alternative response [39,61,66].

Although the fronto-central negativity in our experiments was error-specific, we can be less certain of its sensitivity to the magnitude of that error. A significant effect of expectancy was observed in Experiment 2 when we employed a subtraction analysis designed to minimize the influence of the leading edge of the P3a. A similar analysis in Experiment 1 also was suggestive of an expectancy effect, however this analysis did not reach conventional levels of significance, perhaps because of greater inter-individual variability associated with processing the more complex 'dual-feedback.' The results from Experiment 2, however, are consistent with the findings of previous ERN studies by Holroyd and colleges [39,66] that manipulated the expectancy of negative feedback. By extension, these findings are also compatible with the neural network model of Holroyd and Coles [39], which proposes that the amplitude of the feedback-locked ERN indexes a phasic alteration in mesencephalic dopaminergic input to the ACC that is proportional to the degree to which outcomes are worse than expected. Nonetheless, the relationship between the ERN and the magnitude of error is far from resolved. Other studies suggest that the feedback-locked ERN may be elicited by a fast, automatic detection system simply responding to the absence of reward, prior to the complete evaluation of the magnitude

[4]The specificity of this negative deflection to errors and its equivocal relationship to stimulus probability appears to distinguish it from the N200, a negative waveform with a fronto-central topography that typically demonstrates a similar pattern of results to the P300 component that it immediately precedes (e.g., Ref. [60]). Although previous studies of the feedback-locked ERN have not examined the relationship of the ERN and N200 in detail, at least one study of conflict has isolated a conflict-specific negative deflection in stimulus-locked averages that could be differentiated from the N200 [92]. Furthermore, some studies of the response-locked ERN also suggest that this ERN can be differentiated from the N200 [21,69]. Yet, others have modeled them as generated by a common dipole [87]. Thus, although the relationship between the feedback-locked ERN and N200 has not been definitively resolved, we believe it is unlikely that the ERN-like negativity in the present study is simply an N200.

of that loss or its consequences. For example, Gehring and Willoughby [34] recently observed a feedback-locked medial frontal negativity (MFN) in a gambling task that was modulated only by whether the feedback signaled gain or loss. Similar to the present findings, it was not sensitive to the absolute magnitude of the loss (e.g., 5 vs. 25 US cents), even though there were situations where both response choices would have resulted in loss, and therefore, a lesser loss (e.g., 5) could be considered a 'correct' choice, whereas a greater loss (e.g., 25) could be considered an error. Thus, the feedback-locked ERN in the present studies may have been less sensitive to confidence than the P3a because it was elicited automatically by the penalty implicit in the negative reinforcement stimulus, prior to the conscious appreciation of any further violation in metamemorial expectations [24,25].

Processing of the violation between expected and actual outcome clearly modulated a prominent fronto-central positivity, which may be analogous to the P3a. This deflection peaked 60–75 ms after the ERN, supporting the view that the performance outcome is processed first, followed by the comparison of the outcome with expectations maintained in working memory. Notably, studies in which the response-locked ERN was unrelated to conscious error detection found a significant effect of awareness on the positive deflection that followed (i.e., $P_E$ [65]). Unfortunately, studies in which an expectancy effect was found on the feedback-locked ERN did not examine the P3 [39,66]. In addition, although previous studies using concept generation or paired-associated learning tasks found a feedback-locked positivity that was sensitive to discrepancies between response confidence and outcome [15,40], this positivity appeared to have a more posterior distribution than the P3 that was sensitive to metamemory mismatch in the present studies. Interestingly, although a posterior positivity (P3b) was observed in our studies, it did not appear to be modulated by metamemory mismatch.

The novelty-P3/P3a has been hypothesized to represent activity associated with evaluating and/or orienting to novel events for the purpose of subsequent action [27]. Among the putative generators for this component are neural regions associated with working memory and executive control, including the anterior cingulate cortex (ACC) and prefrontal cortex [2,47], as well as regions involved in the storage of long-term episodic memories, including the hippocampus [48]. The location of these generators suggests that this positivity may index a conflict or mismatch between previous stimulus representations stored in declarative memory and the information currently in working memory. Events eliciting a P3a in the present experiment represented a mismatch not only between the expected and actual outcome of a single trial, but also with the expectations built over the course of the experiment. Subjects' metamemory concerning their response accuracy was very good overall, and thus, metamemory mismatches rarely occurred.

### *4.3. Error correction: the contribution of attention and familiarity*

As in many studies of the response-locked and feedback-locked ERN, the amplitude of this waveform was not predictive of error remediation [31,33,61,65,78]. Although results from others suggest that the ERN may be tied to processes that can serve to inhibit a response that has received negative reinforcement [39,66], it does not appear to directly facilitate processes involved in encoding the correct response into episodic memory.

The P3a demonstrated a relationship to subsequent memory, however this relationship was reliable only for items retested after a relatively short delay. Indeed, the P3a did not differentiate retrieval success at a 1-week delay, when we would expect the effects of emotion-mediated memory consolidation to have been more apparent [38]. This calls into question the hypothesis that the effects of metamemory mismatch were mediated primarily through emotional salience. Rather, these results suggest that this component was an index of the general arousal and level of attention oriented toward the stimulus as a function of its novel or surprising nature. This interpretation would be consistent with the finding that metamemory-mismatch feedback appears to capture more attention, as measured by impairment on a secondary tone-detection task, than does metamemory-match feedback [7]. Increased attention may have provided an additional 'boost' to immediate hypercorrection, as failure to detect the tone also predicted retest accuracy for non-omit errors in that study. If there was such a boost, it seems plausible that it was in the form of a more vivid episodic memory for the trial event, which included the corrective information. 'Remember' responses typically decline with increase in study-test delay, whereas the percentage of 'know' responses stays the same or increases [12,50], suggesting that vivid, self-referential information is more susceptible to decay with time. Yet, because the amplitude of the P3a was not enhanced for items remembered at immediate retest as much as it was attenuated for items forgotten at immediate retest, it is also possible that the temporary memory benefit of metamemory mismatch may be due less to the 'hyper-encoding' of items eliciting metamemory mismatch than to the failure for some low metamemory mismatch items to capture even a basic level of attention necessary to facilitate subsequent semantic processing.

Familiarity had a pervasive effect on error correction that was mirrored by a broad, negative-going waveform following presentation of the correct answer. This waveform was larger for familiar than unfamiliar corrective feedback, and was enhanced further when that information was both familiar and successfully retained, at least up to a delay of one week. Thus, it may index processes associated with Thorndike's [85] second stage of learning, in which associations between pre-existing information in semantic memory are formed or strengthened. Yet, the finding that

presentation of the correct answer also elicited a larger inferior posterior-temporal negativity when the participant had just generated that answer themselves (i.e., correct response) compared to when the participant had generated a different, incorrect answer, suggests that its amplitude is sensitive to the retrieval fluency of an individual item as well. Presumably, recently generated items would carry residual activation that would facilitate retrieval of conceptual information when that same item is provided during feedback.

The relationship of the broad inferior temporal negativity to successful encoding of familiar information, as well as its time course and spatial distribution, suggests that it is similar to the inferior temporal N340 recorded by Mangels and colleagues in a previous study of episodic memory [58]. In that study, the amplitude of the N340 elicited during the encoding of individual words predicted whether that word would be successfully retrieved on a subsequent recognition test, but did not distinguish whether the item would be consciously recollected (i.e., 'remembered') or judged as 'old' on the basis of familiarity (i.e., 'known'). Thus, these results converge with the present findings in suggesting that inferior temporal activity from 300 to 600 ms represents an item-specific conceptual process that can increase subsequent retrieval fluency, although it alone does not appear to provide the contextual and self-referential associative processes that are necessary for a 'remember' response. The distribution of this waveform is also consistent with findings from blood-flow studies demonstrating increased activity within left inferior prefrontal and lateral temporal regions during successful incidental encoding of verbal information (e.g., Refs. [44,46]), as well as studies linking these regions to the storage and retrieval of semantic knowledge (e.g., Refs. [72,91]). Finally, although the inferior temporal negativity in the present studies could be interpreted as indexing either item-specific fluency and/or the strengthening of the question–answer association, future studies in which individual items are made familiar through multiple presentations outside the context of the general information task (e.g., in a lexical decision task) may allow us to disambiguate what aspects are differentially modulated by item-specific and relational processes.

If the inferior temporal negativity indexes a process specific to the encoding of familiar correct answers into long-term memory, then what neurocognitive processes subserve memory for entirely novel information, such as (for most people) remembering that gods in the Norse creation myth reside in 'Asgaard'? The answer is not clear from the present data. Although unfamiliar correct answers were more poorly remembered, particularly at the 1-week delay, retrieval success at immediate retest was a moderate 0.22 in Experiment 1 and 0.31 in Experiment 2. Yet, the waveforms for remembered and forgotten unfamiliar items appeared identical across the scalp (see Fig. 4). It is possible that a null effect resulted in part from variability due to the relatively low trial counts of remembered familiar items. In addition, we cannot rule out the possibility that the act of rating familiarity itself contributed to these low trial counts by biasing attention and elaborative processes toward familiar items, enhancing both the amplitude of the inferior temporal negativity and the likelihood that these familiar items would be remembered. Yet, poor memory for unfamiliar answers also suggests that the system underlying episodic encoding is fundamentally more attuned to the novel combination of semantically familiar elements (e.g., a familiar word in an usual context) than entirely novel semantic information that might initially bear greater resemblance to a pseudoword (i.e., Asgaard). Although recent studies have shown that many of the same neural regions that are sensitive to novelty are also correlated with episodic encoding success (e.g., Refs. [17,46,83,84,86,89]), these studies defined novelty by the number of repetitions of an item with a pre-existing representation in semantic memory or the novel juxtaposition of familiar elements. In contrast, words without pre-existing semantic representations are most likely encoded through phonological or lexical representations, and may require extensive repetition in order to become integrated into semantic memory. To the extent that this repetition would have occurred outside of the 1.5 s epoch following presentation of the word, it would not have been apparent in our averages.

### 4.4. Conclusion

We often say that we learn best from our mistakes. The present studies provides insight into the sequence of cognitive and neural events underlying this learning process when the information to be learned is semantic knowledge. Following feedback about the accuracy of one's response to a trivia question, subjects appear to first engage in the detection of the negative outcome, as indexed by a fronto-central ERN-like wave, possibly through a fast, automatic system gated by the presence of negative reinforcement stimuli that is also sensitive to the degree to which that negative outcome is worse than expected. They then orient attention to this information in proportion to the degree to which the outcome deviates from expectation, as indexed by the amplitude of a fronto-central positivity, putatively the P3a—an orienting response that also occurs in the presence of unexpected positive outcomes. Correction of high-confidence errors appeared to benefit from this orienting response to metamemory mismatch, given that the amplitude of the P3a was also related to error correction, at least when performance was retested at a relatively short delay. In addition, these high-confidence errors benefited from the greater likelihood that their correct answers would be familiar. Indeed, the familiarity of corrective information facilitated error correction at both immediate and 1-week delayed retest, an effect that was mirrored by a broad inferior temporal negativity that may represent either item-

specific fluency and/or the strengthening of the question–answer association.

## Acknowledgements

This research was supported by NIH grant R21MH066129 and a grant from the W.M. Keck Foundation. We thank Paul Ferber and Christopher Romero for programming and graphic support, and Amelia Kaplan, Jisun Lee, Cecilia Lipira, Juliana Oak, Derek Nagy, and Daniel Wetmore for assistance in running subjects and analyzing data. Aspects of this data were presented as a poster at the 2000 meeting of the Society for Psychophysiological Research, San Diego, CA, for which B. Butterfield received the Tursky Award, and the 2002 Cognitive Neuroscience Society Meeting, San Francisco, CA.

## Appendix A

The algorithm for question selection in Experiment 2 chose the question for each trial according to the level of current performance: if the participant's current proportion of correct trials was 0.40 (rounded to two decimal places), or if it was the first trial of the experiment, a question was chosen randomly from the pool of remaining questions and this question was presented to the participant. However, if the current proportion correct was less than 0.40, a random question was selected and a random number between 0 and 1 was generated. This question was presented if the normative ease of the question (proportion of pilot subjects who answered it correctly) was greater than the random number. If it was not, a different question was selected and a different random number was generated. This procedure was iterated until a question was found that had a normative rating higher than the random number, for a maximum of four iterations. The net effect of this process was that, when performance was less than 0.40, easier questions were more likely to be presented, although any remaining questions in the pool also had a chance of being presented. The same basic process was used if the current proportion correct was greater than 0.40, with the difference that the selected question was presented if its normative ease was less than the random number. The net effect of this process was that, when performance was greater than 0.40, difficult questions were more likely to be presented, although, again, any remaining questions also had a chance of being presented.

## References

[1] J.R. Anderson, G.H. Bower, Recognition and retrieval processes in free recall, Psychol. Rev. 79 (1972) 97–123.
[2] P. Baudena, E. Halgren, G. Heit, J.M. Clarke, Intracerebral potentials to rare target and distractor auditory and visual stimuli. III. Frontal cortex, Electroencephaolgr. Clin. Neurophysiol. 94 (1995) 251–264.
[3] A.S. Benjamin, R.A. Bjork, B.L. Schwartz, The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index, J. Exp. Psychol. Gen. 127 (1998) 55–68.
[4] P. Berg, M. Scherg, A multiple source approach to the correction of eye artifacts, Electroencephaolgr. Clin. Neurophysiol. 90 (1994) 97–105.
[5] M.M. Botvinick, T.S. Braver, D.M. Barch, C.S. Carter, J.D. Cohen, Conflict monitoring and cognitive control, Psychol. Rev. 108 (2001) 624–652.
[6] B. Butterfield, J. Metcalfe, Errors committed with high confidence are hypercorrected, J. Exp. Psychol. Learn. Mem. Cogn. 27 (2001) 1491–1494.
[7] B. Butterfield, J. Metcalfe, The role of attention in the hypercorrection effect, in: 43rd Annual Meeting of the Psychonomic Society, Kansas City, MO, 2002.
[8] L. Cahill, J.L. McGaugh, Mechanisms of emotional arousal and lasting declarative memory, Trends Neurosci. 21 (1998) 294–299.
[9] T. Canli, Z. Zhao, J. Brewer, J.D.E. Gabrieli, L. Cahill, Event-related activation in the human amygdala associates with later memory for individual emotional experience, J. Neurosci. 20 (2000) 1–5.
[10] M.G.H. Coles, M.K. Scheffers, C.B. Holroyd, Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing, Biol. Psychol. 56 (2001) 173–189.
[11] M. Comerchero, J. Polich, P3a, perceptual distinctiveness, and stimulus modality, Brain Res. Cogn. Brain Res. 7 (2000) 41–48.
[12] M.A. Conway, J.M. Gardiner, T.J. Perfect, S.J. Anderson, G.M. Cohen, Changes in memory awareness during learning: the acquisition of knowledge by psychology undergraduates, J. Exp. Psychol. Gen. 126 (1997) 393–413.
[13] E. Courchesne, S.A. Hillyard, R. Galambos, Stimulus novelty, task relevance, and the visual evoked potential in man, Electroencephaolgr. Clin. Neurophysiol. 39 (1975) 131–143.
[14] P.L. Davies, S.J. Segalowitz, J. Dywan, P.E. Pailing, Error-negativity and positivity as they related to other ERP indices of attentional control and stimulus processing, Biol. Psychol. 56 (2001) 191–206.
[15] J.H. DeSwart, A. Kok, E.A. Das-Small, P300 and uncertainty reduction in a concept-identification task, Psychophysiology 18 (1981) 619–629.
[16] C.S. Dodson, W. Koutstaal, D.L. Schacter, Escape from illusion: reducing false memories, Trends Cogn. Sci. 4 (2000) 391–397.
[17] R.J. Dolan, P.C. Fletcher, Dissociating prefrontal and hippocampal function in episodic memory encoding, Nature 388 (1997) 582–585.
[18] E. Donchin, M.G.H. Coles, Is the P300 component a manifestation of context updating?, Behav. Brain Sci. 11 (1988) 357–474.
[19] C.E. Elger, T. Grunwald, K. Lehnertz, M. Kutas, C. Helmstaedter, A. Brockhaus, D. VanRoost, H.J. Heinze, Human temporal lobe potentials in verbal learning and memory processes, Neuropsychologia 35 (1997).
[20] M. Fabiani, E. Donchin, Encoding processes and memory organization: a model of the von Restorff effect, J. Exp. Psychol. Learn. Mem. Cogn. 21 (1995) 224–240.
[21] M. Falkenstein, J. Hohnsbein, J. Hoormann, ERP components in Go/Nogo tasks and their relation to inhibition, Acta Psychol. 101 (1999) 267–291.
[22] M. Falkenstein, J. Hoormann, S. Christ, L. Blanke, Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks, Electroencephaolgr. Clin. Neurophysiol. 78 (1991) 447–455.
[23] M. Falkenstein, J. Hoormann, S. Christ, J. Hohnsbein, ERP components on reaction errors and their functional significance: a tutorial, Biol. Psychol. 51 (2000) 87–107.
[24] D. Fernandez-Duque, J.A. Baird, M.I. Posner, Awareness and metacognition, Conscious Cogn. 9 (2000) 324–326.

[25] D. Fernandez-Duque, J.A. Baird, M.I. Posner, Executive attention and metacognitive regulation, Conscious Cogn. 9 (2000) 288–307.
[26] C. Fiorillo, P. Tobler, W. Schultz, Discrete coding of reward probability and uncertainty by dopamine neurons, Science 299 (2003) 1898–1902.
[27] D. Friedman, Y.M. Cycowicz, H. Gaeta, The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty, Neurosci. Biobehav. Rev. 25 (2001) 355–373.
[28] J.D.E. Gabrieli, J.E. Desmond, J.B. Demb, A.D. Wagner, M.V. Stone, C.J. Vaidya, G.H. Glover, Functional magnetic resonance imaging of semantic memory processes in the frontal lobes, Psychol. Sci. (1996) 278–283.
[29] H. Gaeta, D. Friedman, G. Hunt, Stimulus characteristics and task category dissociate the anterior and posterior aspects of the novelty P3, Psychophysiology 40 (2003) 198–208.
[30] D.A. Gallo, K.B. McDermott, J.M. Percer, H.L.I. Roediger, Modality effects in false recall and false recognition, J. Exp. Psychol. Learn. Mem. Cogn. 27 (2001) 339–353.
[31] W.J. Gehring, D.E. Fencsik, Functions of the medial frontal cortex in the processing of conflict and errors, J. Neurosci. 21 (2001) 9430–9437.
[32] W.J. Gehring, B. Goss, M.G.H. Coles, D.E. Meyer, E. Donchin, A neural system for error detection and compensation, Psychol. Sci. 4 (1993) 385–390.
[33] W.J. Gehring, R.T. Knight, Prefrontal–cingulate interactions in action monitoring, Nat. Neurosci. 3 (2000) 516–520.
[34] W.J. Gehring, A.R. Willoughby, The medial frontal cortex and the rapid processing of monetary gains and losses, Science 295 (2002) 2279–2282.
[35] G. Gillund, R.M. Shiffrin, A retrieval model for both recognition and recall, Psychol. Rev. 91 (1984) 1–67.
[36] A. Goldstein, K.M. Spencer, E. Donchin, The influence of stimulus deviance and novelty on the P300 and Novelty P3, Psychophysiology 39 (2002) 781–790.
[37] T. Grunwald, H.J. Heinze, C.E. Elger, Verbal novelty detection within the human hippocampus proper, Proc. Natl. Acad. Sci. USA 95 (1998) 3193–3197.
[38] S.B. Hamann, T.D. Ely, S.T. Grafton, C.D. Kilts, Amygdala activity related to enhanced memory for pleasant and aversive stimuli, Nat. Neurosci. 2 (1999) 289–293.
[39] C.B. Holroyd, M.G.H. Coles, The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity, Psychol. Rev. 109 (2002) 679–709.
[40] R.L. Horst, R.J. Johnson, E. Donchin, Event-related brain potentials and subjective probability in a learning task, Mem. Cogn. 8 (1980) 476–488.
[41] J.R. Jennings, C.C. Wood, The e-adjustment procedure for repeated-measure analyses of variance, Psychophysiology 13 (1976) 277–278.
[42] W.A. Johnston, V.J. Dark, L.L. Jacoby, Perceptual fluency and recognition judgments, J. Exp. Psychol. Learn. Mem. Cogn. 11 (1985) 3–11.
[43] J. Kaiser, R. Barker, C. Haenschel, T. Baldeweg, J.H. Grunzelier, Hypnosis and event-related potential correlates of error processing in a stroop-type paradigm: a test of the frontal hypothesis, Int. J. Psychophysiol. 27 (1997) 215–222.
[44] S. Kapur, F.I.M. Craik, E. Tulving, A.A. Wilson, S. Houle, E. Tulving, Neuroanatomical correlates of encoding in episodic memory: levels of processing effect, Proc. Natl. Acad. Sci. USA 91 (1994) 2008–2011.
[45] C.M. Kelley, S.D. Lindsay, Remembering mistaken for knowing: ease of retrieval as a basis for confidence in answers to general knowledge questions, J. Mem. Lang. 34 (1993) 1–24.
[46] B.A. Kirchhoff, A.D. Wagner, A. Maril, C.E. Stern, Prefrontal-temporal circuitry for episodic encoding and subsequent memory, J. Neurosci. 20 (2000) 6173–6180.
[47] R.T. Knight, Decreased response to novel stimuli after prefrontal lesions in man, Electroencephaolgr. Clin. Neurophysiol. 59 (1984) 9–20.
[48] R.T. Knight, Contribution of the human hippocampal region to novelty detection, Nature 383 (1996) 256–259.
[49] R.T. Knight, D. Scabini, Anatomic bases of event-related potentials and their relationship to novelty detection in humans, J. Clin. Neurophysiol. 15 (1998) 3–13.
[50] B.J. Knowlton, L.R. Squire, Remembering and knowing: two different expressions of declarative memory, J. Exp. Psychol. Learn. Mem. Cogn. 21 (1995) 699–710.
[51] A. Kok, On the utility of the P3 amplitude as a measure of processing capacity, Psychophysiology 38 (2001) 557–577.
[52] A. Koriat, M. Goldsmith, A. Pansky, Toward a psychology of memory accuracy, Annu. Rev. Psychol. 51 (2000) 481–537.
[53] M. Kutas, Views on how the electrical activity that the brain generates reflects the functions of difference language structures, Psychophysiology 34 (1997) 383–398.
[54] M. Kutas, K.D. Federmeier, Electrophysiology reveals semantic memory use in language comprehension, Trends Cogn. Sci. 4 (2000) 463–470.
[55] M. Kutas, S. Hillyard, Reading senseless sentences: brain potentials reflect semantic incongruity, Science 207 (1980) 203–205.
[56] E.F. Loftus, D.C. Polage, Repressed memories. When are they real? How are they false?, Psychiatr. Clin. North Am. 22 (1999) 61–70.
[57] P. Luu, P. Collins, D.M. Tucker, Mood, personality and self-monitoring: negative affect and emotionality in relation to frontal lobe mechanisms of error monitoring, J. Exp. Psychol. Gen. 129 (2000) 43–60.
[58] J.A. Mangels, T.W. Picton, F.I.M. Craik, Attention and successful episodic encoding: an event-related potential study, Brain Res. Cogn. Brain Res. 11 (2001) 77–95.
[59] J. Metcalfe, Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony, J. Exp. Psychol. Gen. 119 (1990) 145–160.
[60] H.J. Michalewski, D.K. Prasher, A. Starr, Latency variability and temporal interrelationships of the auditory event-related potentials (N1, P2, N2, and P3) in normal subjects, Electroencephaolgr. Clin. Neurophysiol. 65 (1986) 59–71.
[61] W.H.R. Miltner, C.H. Braun, M.G.H. Coles, Event-related potentials following incorrect feedback in a time-estimation task: evidence for a 'generic' neural system for error detection, J. Cogn. Neurosci. 9 (1997) 788–798.
[62] T.O. Nelson, L. Narens, Norms of 300 general-information questions: accuracy of recall, latency of recall, and feeling-of-knowing ratings, J. Verb. Learn. Verb. Behav. 19 (1980) 338–368.
[63] T.O. Nelson, L. Narens, Why investigate metacognition, in: J. Metcalfe, A.P. Shimamura (Eds.), Metacognition: Knowing About Knowing, MIT Press, Cambridge, MA, 1994.
[64] H.J. Neville, M. Kutas, G. Chesney, A.L. Schmidt, Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words, J. Mem. Lang. 25 (1986) 75–92.
[65] S. Nieuwenhuis, K.R. Ridderinkhof, J. Blom, G.P.H. Band, A. Kok, Error-related brain potentials are differently related to awareness of response errors: evidence from an antisaccade task, Psychophysiology 38 (2001) 752–760.
[66] S. Nieuwenhuis, K.R. Ridderinkhof, D. Talsma, M.G.H. Coles, C.B. Holroyd, A. Kok, M.W. van der Molen, A computational account of altered error processing in older age: dopamine and the error-related negativity, Cogn. Affect. Behav. Neurosci. 2 (2002) 19–36.
[67] A.C. Nobre, G. McCarthy, Language-related ERPs: scalp distributions and modulation by word type and semantic priming, J. Cogn. Neurosci. 6 (1994) 233–255.
[68] A.C. Nobre, G. McCarthy, Language-related field potentials in the anterior-medial temporal lobe. II. Effects of word type and semantic priming, J. Neurosci. 15 (1995) 1090–1098.
[69] P.E. Pailing, S.J. Segalowitz, P.L. Davies, Speed of responding and the likelihood of error-like activity in correct trial ERPs, Psychophysiology 37 (2000) S76.

[70] T.W. Picton, The P300 wave of the human event-related potential, J. Clin. Neurophysiol. 9 (1992) 456–479.
[71] T.W. Picton, S. Bentin, P. Berg, E. Donchin, S. Hillyard, R.J. Johnson, G. Miller, W. Ritter, D. Ruchkin, M. Rugg, M. Taylor, Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria, Psychophysiology 37 (2000) 127–152.
[72] C.J. Price, The anatomy of language: contributions from functional neuroimaging, J. Anat. 197 (Pt 3) (2000) 335–359.
[73] P.M.A. Rabbit, B. Rodgers, What does a man do after he makes an error? An analysis of response programming, Q. J. Exp. Psychol. 29 (1977) 727–743.
[74] R.A. Rescorla, A.R. Wagner, A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement, in: A.H. Black, W.F. Prokasy (Eds.), Classical Conditioning II: Current Research and Theory, Appleton–Century–Crofts, New York, 1972, pp. 64–99.
[75] M. Ruchsow, J. Grothe, M. Spitzer, M. Kiefer, Human anterior cingulate cortex is activated by negative feedback: evidence from event-related potentials in a guessing task, Neurosci. Lett. 325 (2002) 203–206.
[76] M.D. Rugg, M.G.H. Coles, Electrophysiology of Mind: Event-related Potentials and Cognition, Oxford University Press, Oxford, 1995.
[77] M.K. Scheffers, M.G.H. Coles, Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors, J. Exp. Psychol. Hum. Percept. Perform. 26 (2000) 141–151.
[78] M.K. Scheffers, M.G.H. Coles, P. Bernstein, W.J. Gehring, E. Donchin, Event-related brain potentials and error-related processing: an analysis of incorrect responses to go and no-go stimuli, Psychophysiology 33 (1996) 42–53.
[79] W. Schultz, Getting formal with dopamine and reward, Neuron 36 (2002) 241–263.
[80] R.F. Simons, F.K. Graham, M.A. Miles, X. Chen, On the relationship of the P3a and the Novelty-P3, Biol. Psychol. 56 (2001) 207–218.
[81] K.M. Spencer, J. Dien, E. Donchin, A componential analysis of the ERP elicited by novel events using a dense electrode array, Psychophysiology 36 (1999) 409–414.
[82] K.M. Spencer, J. Dien, E. Donchin, Spatiotemporal analysis of the late ERP responses to deviant stimuli, Psychophysiology 38 (2001) 343–358.
[83] R.A. Sperling, J.F. Bates, A.J. Cocchiarella, D.L. Schacter, B.R. Rosen, M.S. Albert, Encoding novel face–name associations, Hum. Brain Mapp. 14 (2001) 129–139.
[84] C.E. Stern, S. Corkin, R.G. Gonzalez, A.R. Guimaraes, J.R. Baker, P.J. Jennings, C.A. Carr, R.M. Sugiura, V. Vedantham, B.R. Rosen, The hippocampal formation participates in novel picture encoding: evidence from functional magnetic imaging, Proc. Natl. Acad. Sci. USA 93 (1996) 8660–8665.
[85] E.L. Thorndike, The Fundamentals of Learning, Bureau of Publications, Teachers College, New York, 1932, pp. 638.
[86] E. Tulving, H.J. Markowitsch, F.I.M. Craik, R. Habib, S. Houle, Novelty and familiarity activations in PET studies of memory encoding and retrieval, Cereb. Cortex 71–79 (1996).
[87] V. van Veen, C.S. Carter, The timing of action-monitoring processing the anterior cingulate cortex, J. Cogn. Neurosci. 14 (2002) 593–602.
[88] F. Vidal, T. Hasbroucq, J. Grapperon, M. Bonnet, Is the 'error negativity' specific to errors?, Biol. Psychol. 51 (2000) 109–128.
[89] A.D. Wagner, W. Koutstaal, A. Maril, D.L. Schacter, R.L. Buckner, Task-specific repetition priming in left inferior prefrontal cortex, Cereb. Cortex 10 (2000) 1176–1184.
[90] A.D. Wagner, W. Koutstaal, D.L. Schacter, When encoding yields remembering: insights from event-related neuroimaging, Philos. Trans. R. Soc. London B Biol. Sci. 354 (1999) 1307–1324.
[91] A.D. Wagner, E.J. Pare-Blagoev, J. Clark, R.A. Poldrack, Recovering meaning: left prefrontal cortex guides controlled semantic retrieval, Neuron 31 (2001) 329–338.
[92] Y. Wang, J. Kong, X. Tang, D. Zhuang, S. Li, Event-related potential N270 is elicited by mental conflict processing in human brain, Neurosci. Lett. 293 (2000) 17–20.
[93] C.D. Woody, Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals, Med. Biol. Eng. 5 (1967) 529–553.