Neural Correlates of People's Hypercorrection of Their False Beliefs

Janet Metcalfe, Brady Butterfield, Christian Habeck, and Yaakov Stern

Abstract

■ Despite the intuition that strongly held beliefs are particularly difficult to change, the data on error correction indicate that general information errors that people commit with a high degree of belief are especially easy to correct. This finding is called the *hypercorrection effect*. The hypothesis was tested that the reason for hypercorrection stems from enhanced attention and encoding that results from a metacognitive mismatch between the person's confidence in their responses and the true answer. This experiment, which is the first to use imaging to investigate the hypercorrection effect, provided support for this hypothesis, showing that both metacognitive mismatch conditions—that in which high confidence accompanies a wrong answer and that in which low confidence accompanies a correct answer—revealed anterior cingulate and medial frontal gyrus activations. Only in the high confidence error condition, however, was an error that conflicted with the true answer mentally present. And only the high confidence error condition yielded activations in the right TPJ and the right dorsolateral pFC. These activations suggested that, during the correction process after error commission, people (1) were entertaining both the false belief as well as the true belief (as in theory of mind tasks, which also manifest the right TPJ activation) and (2) may have been suppressing the unwanted, incorrect information that they had, themselves, produced (as in think/no-think tasks, which also manifest dorsolateral pFC activation). These error-specific processes as well as enhanced attention because of metacognitive mismatch appear to be implicated. ■

INTRODUCTION

Although the production of errors is a human shortcoming that has been subject to considerable study, the processes involved in how people correct those errors have only recently begun to receive intensive scrutiny. There is no doubt that understanding the processes of error correction is central to understanding human cognition insofar as the correction of mistakes and the updating of our knowledge are necessary for learning in all domains. People need to correct errors to be in a position to make realistic and appropriate decisions, whether a student correcting incorrect information for a future test, a physician correcting a diagnosis, or a teacher, shopkeeper, business person, police officer, judge, or president keeping their knowledge correct and current. Indeed, the progress of scientific investigation itself can be seen as an ongoing process of error generation (of wrong hypotheses) and error correction as we refine our knowledge.

There are now many studies showing that, unless the individual is given feedback concerning the correct answer rather than only a signal that an error has been made, later accurate responding is unlikely (Metcalfe & Finn, 2010; Pashler, Cepeda, Wixted, & Rohrer, 2005). In addition, the person's confidence or degree of belief in the original error has consequences for both the processes and probability of error correction. It is this relation between confidence and error correction that is the focus of this study.

There is considerable evidence that people's metacognitions concerning their responses are generally accurate (Metcalfe & Finn, 2008; Koriat, Goldsmith, & Pansky, 2000; Koriat, 1993, 1997; Murdock, 1974). When people express a high degree of confidence in a response, they are likely to be correct. Low confidence is associated with incorrect responses. Presumably, the reason for the high correlation between people's confidence in their answers and the correctness of those particular answers is that high confidence is associated with strong and easily accessible memories and beliefs (see Finn & Metcalfe, 2010; Koriat, 1993, 2008, for a model of this relation). Responses given with low confidence are likely to be either guesses or weaker and less-entrenched beliefs. Given that high confidence self-generated responses are strong and interference resistant, they should also be particularly difficult to alter. Many perspectives from interference theory (Barnes & Underwood, 1959) and PDP (parallel distributed processing) and computational models of memory (e.g., Rumelhart & McClelland, 1987) and many experiments in the misleading information (e.g., Loftus & Hoffman, 1989) and false memory paradigms (e.g., Fazio & Marsh, 2009; Roediger & McDermott,

Columbia University

© 2012 Massachusetts Institute of Technology

1995) converge on the conclusion that high confidence is associated with strong and accessible responses.

It follows that if a high confidence response is erroneous, it, too—like correct high confidence responses—should be strong and resistant to being overwritten. In stark contrast to this prediction, however, it has been found (Butterfield & Metcalfe, 2001; see Kulhavy, Yekovich, & Dyer, 1976) that high confidence errors are easier to correct than low confidence errors. This surprising phenomenon is called the "high confidence error hypercorrection effect" or, sometimes, the "hypercorrection effect." In typical experiments showing this effect, people are given a series of general information questions. After each question, they give their answers, followed by their confidence in the correctness of their answers. They are then given the correct answer as feedback. At some later time, participants are tested for the correct answers. The now-standard finding is that errors that were expressed with higher confidence are more likely to be corrected on this final test than are errors that were committed with lower confidence. The original hypercorrection findings have now been replicated many times (Metcalfe & Finn, 2011, in press; Butler, Fazio, & Marsh, 2010; Fazio & Marsh, 2009, 2010; Sitzman & Rhodes, 2010; Butterfield & Metcalfe, 2001, 2006; Butterfield & Mangels, 2003; Kulhavy & Stock, 1989; Kulhavy et al., 1976), and a dominant hypothesis concerning the locus of the phenomenon has emerged.

The dominant explanation—an explanation that we will both investigate and extend in the present imaging experiment—posits that the hypercorrection effect occurs because there is enhanced attention to the correct responses for high confidence errors because of metacognitive mismatch. When people experience the contrast of their beliefs in their answers with the knowledge that that answer is incorrect, the dissonance results in heightened attention being paid to the corrective feedback. To give an example: When people say with high confidence that the capital of Canada is Toronto, only to find that it is actually Ottawa, they are surprised by their high confidence errors. The heightened attention because of the metacognitive mismatch results in hyperencoding of the correction and enhanced memory for it.

Note that, in the error correction paradigm, there are two conditions that exhibit metacognitive mismatch. One is the case in which the person makes a high confidence error and finds out that he or she is wrong. The other is the case in which the person makes a low confidence correct response—essentially, a guess out of the blue—and finds out that he or she is correct.

Three lines of research support the metacognitive mismatch hypothesis as a locus of hypercorrection of high confidence errors. First, Butterfield and Metcalfe (2006) tested the hypothesis directly by having participants engage in a simultaneous attention-demanding tone-detection task while they were answering questions and receiving corrective feedback. They were interested in both high confidence errors and low confidence corrects. Although these latter responses are not errors, Butler, Karpicke, and Roediger (2008) have shown that, without feedback, the participant is very likely to produce an incorrect response later. Butterfield and Metcalfe's (2006) rationale was that, if the hypercorrection effect was because of enhanced attention to the surprising response due to metacognitive mismatch, then participants would be likely to miss hearing a faint tone that required attention to detect, if it was presented during the (visual) feedback in both the high confidence error and low confidence correct metacognitive mismatch conditions. This symmetrical and selective attentional effect was what they found.

Fazio and Marsh (2009), too, proposed that the hypercorrection effect obtains because people invest extra attention in those particular corrections because of surprise at being wrong or what we call here metacognitive mismatch. They argued that if the effect was because of attention then the background setting of the to-be-remembered correct responses would be better remembered for the high than the low confidence errors. And, indeed, they found such enhanced memory for the context. And finally, in an ERP investigation of the hypercorrection effect, Butterfield and Mangels (2003) examined the scalp voltage elicited by the feedback to correct and erroneous responses. They found a fronto-central P3 component that was associated with both metacognitive mismatch conditions. The P3 is a positive component peaking at around 300 msec after stimulus onset that is classically elicited by rare/novel events such as in an "oddball" paradigm and is thought to indicate enhanced attention to the unexpected stimulus (Picton, 1992). It has been found to be associated with memory enhancement both in Butterfield and Mangel's hypercorrection study and in general (Paller, McCarthy, & Wood, 1988). The amplitude of the P3 component in Butterfield and Mangels' (2003) experiments was graded by the extent of metamemory mismatch: It was greater to the feedback to high confidence errors than to medium confidence errors than to low confidence errors. It was also greater to the feedback to low confidence correct responses than to medium confidence correct responses than to high confidence correct responses.

These three lines of research suggest that an attentionally related process probably implicating anterior cingulate and related frontal areas and attributable to the metacognitive mismatch should be found. Such an activation should be associated with both metacognitive mismatch conditions: people's reactions to the feedback to high confidence errors and to the feedback to low confidence corrects. To further investigate the metacognitive mismatch hypothesis, the present experiment, which is the first imaging study directed at the hypercorrection effect, used fMRI to determine the neural correlates of both metamemory mismatch conditions during feedback in the scanner. Although our primary interest centered on the high confidence errors, the low confidence corrects presented the mirror image of metacognitive mismatch

in the prior behavioral studies. Thus, to the extent that expectation violation was the underlying factor giving rise to the hypercorrection effect, we expected the brain correlates in the two conditions to be the same (as in the experiments of Butterfield & Mangels, 2003, and Butterfield & Metcalfe, 2006, in which these two conditions had shown symmetrical effects; see Hunt, 2009).

Although not disputing the importance of attentional processes in hypercorrection, other factors also appear to be involved. In the high confidence error condition, there is a conflict between what one believes to be true and what one is told is true. The relation of the errors to the correct answers has been shown to differ between high and low confidence errors (Metcalfe & Finn, 2011, in press), with high confidence errors being more related to the correct answer than low confidence errors. Huelser and Metcalfe (2011) have shown that error generation, per se, facilitates later correct recall and that the benefit is greater when the errors are related as compared with unrelated to the target. Of course, in the low confidence correct metacognitive mismatch condition, there is no error present at all. It follows that, although the low confidence correct metacognitive mismatch condition should be similar to the high confidence error condition in terms of the surprise/attention reaction, it should be different because no error is present in the low confidence correct case. In contrast, a mistaken belief, along with the correct answer, exists in the high confidence error case. In the condition in which a high confidence error was committed, then, two conflicting pieces of information are present: the answer that the participants produced and affirmed that they believed and the conflicting answer that they have been told is correct. When no error was committed, only the correct response produced by the participant is present. A differential brain response, then, should be observable depending upon whether the erroneous belief was present as compared with when there was no erroneous belief. In the high confidence error case (as contrasted to the low confidence correct case), we would expect to see activation in areas associated with consideration and coordination of both incorrect and correct beliefs. Although other factors may also be important in theory of mind (TOM) studies, the need for the participant to consider and coordinate true and false beliefs in those studies suggests that there might be overlapping areas of activation between the TOM and the error correction paradigms.

Furthermore, to remember the correct response when an error has been committed, the learner needs not only to entertain the correct and incorrect beliefs but also to discriminate between the two and/or to suppress the error. Neither discrimination nor suppression is necessary in the low confidence correct condition, of course. Thus, brain activation associated with discrimination/suppression of an erroneous response would also be expected to be selective to the error-correction metacognitive mismatch condition and was not expected in the correct response condition, regardless of confidence rating. Saxe (in press) has suggested that this process is related to dorsolateral pFC (DLPFC) activation. Anderson and Green (2001) have shown that DLPFC activation is associated with inhibiting mental activity in a think/no-think task, in which participants are first asked to learn a cue target pair but then later are asked to inhibit recall of the target in the "no-think" condition. Later, their recall of the target that had been subjected to the no-think condition is impaired (Anderson & Levy, 2006; Anderson & Green, 2001; see also Nee & Jonides, 2008, for the relation of interference control to DLPFC activation). Now, it has not been determined whether the process involved in error correction necessarily involves suppression, of the sort that Anderson and colleagues have been exploring, although it is plausible that it might. If so, the DLPFC would be a likely area that would exhibit activation.

An alternative possibility, though, is that the person might use the retrieved error to mediate retrieval of the targeted correct response. To do so effectively, however, it would be essential that he or she minimally be able to discriminate the correct from the incorrect response. Response discrimination has also been found to implicate the DLPFC (e.g., Bunge, Hazeltine, Scanlon, Rosen, & Gabrieli, 2002; MacDonald, Cohen, Stenger, & Carter, 2000). Thus, DLPFC activation would be expected to occur in the error correction condition regardless of whether the errors were suppressed to allow later correct responding or were remembered and used as (discriminated) mediators to enhance later correct responding.

The hypotheses in this experiment—which is the first imaging experiment to investigate the hypercorrection effect—were that, following feedback to both of the metacognitive mismatch conditions, we would observe fronto-centro activations associated with increased attention based on expectation violation. We expected this activation to be most prominent in and about the anterior cingulate. We also expected that there would be particular activations associated with coordinating two sources of information—the high confidence errors and the correct answers. This coordination of true and false beliefs would be associated with the high confidence error condition but not with the low confidence correct condition.

METHODS

Participants

Fifteen participants (10 women and 5 men, mean age = 22 years) were included in this experiment. The data from one participant were excluded, before being analyzed, because of a failure to demonstrate significant visual cortex activity to the question–response–confidence cue. This activation was used as a "screen" to check for problems in image acquisition. The reason for the failure, in the case of this participant, is unknown. Participants were all right-handed native speakers of English, had normal or corrected-

to-normal vision, and had no history of neurological or psychological disorders. Before testing, participants were screened for their knowledge of general information using a pretest consisting of five easy, five medium, and five difficult questions that were not repeated during the actual test. They were excluded from the experiment if they correctly answered fewer than five or more than 12 questions on this pretest. All participants provided informed consent and were compensated at a rate of $12/hour for their participation.

Materials

The stimuli were 599 general information questions from a variety of knowledge domains. All questions had answers that were familiar to at least 95% of previous participants, and all answers were single words of three to eight letters in length. This question set included questions used in previous studies as well as questions added from Internet sites.

Procedure

Importing the hypercorrection paradigm into fMRI presented several challenges. In previous studies, participants responded to a general information question and then immediately received feedback about response accuracy. This was not possible in the fMRI setting because (1) participants would have had difficulty lying still in a scanner for the multiple hours this procedure entailed and (2) responding to general information questions (by typing or speaking the answers) requires movement that would have created large artifacts. For these reasons, the response acquisition and the feedback presentation were split into separate phases. In the test phase, participants gave responses and rated their confidence in the accuracy of their responses to general information questions while outside the scanner. In the subsequent feedback phase, participants received feedback for a selected set of questions while in the scanner. Finally, in the retest phase, participants took a retest over the selected set of questions outside the scanner.

The purpose of the test phase was to collect 20 trials for each of the seven trial types of interest (response omissions, low/medium/high confidence errors, and low/medium/high confidence corrects). General information 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." For all responses except omits, participants then rated their confidence in their responses on a scale ranging from −50 (*sure incorrect*) to 0 (*not sure if correct or incorrect*) to 50 (*sure correct*). Participants were encouraged to use the entire scale. For the tabulation of trial types of interest, the scale of confidence was split into low confidence (<-17, i.e., the lowest 33 points), medium confidence (between −17 and 17, inclusive, i.e., the middle 35 points), and high confidence (>17, i.e., the highest 33 points) categories. The program used a letter-matching algorithm to score the response as a match (75/100 or greater), nonmatch (less than 70/100), or borderline (greater than 70/100 and less than 75/100), but participants were not given feedback in the first phase. This procedure continued until a minimum of 20 trials of each of the seven trial types had been collected (responses of borderline accuracy were discarded). If the participant answered all 599 questions before fulfilling the trial count requirement, the program shifted the edges of the confidence category bins until 20 trials of each type were present.[1]

After the participant was done with the test phase, the program selected trials for feedback presentation. This selection was accomplished pseudorandomly to fulfill two goals. The first goal was to obtain the highest confidence errors and the lowest confidence corrects to maximize the sensitivity of the design to metamemory mismatch. The second goal was to match errors and corrects on confidence. Thus, the high confidence corrects were not, on average, of higher confidence than the high confidence errors. This was essential because, in the feedback phase, participants were cued with their previous responses and response confidence before given feedback. Matching errors and corrects on confidence minimized the possibility that participants might, based on the confidence with which it was endorsed, guess the accuracy of a given trial before the feedback was presented. To this end, the 20 highest confidence errors were selected for presentation in the scanner. The 20 high confidence corrects were chosen by matching their confidence levels, trial-by-trial, to those of the 20 highest confidence errors. Thus, "high" confidence for high confidence errors and high confidence corrects was as equivalent as possible. Similarly, the 20 lowest confidence correct responses were chosen, and the 20 low confidence errors were chosen based on a match with these low confidence corrects. Medium confidence errors and corrects were also matched in this fashion.

In the feedback phase of the experiment, the participant was given feedback in the scanner for the 140 trials selected from the test phase. Trials were presented in five blocks, with trials randomly assigned to blocks so that four trials of each type (high, medium, and low confidence corrects and errors and response omissions) were randomly mixed within each block. Twenty "blank" intervals, lasting 4 sec each, were also presented in each block. These were used as the baseline for the "screen" test noted above, which was used to check that there were no problems with image acquisition. Participants were instructed to simply pay attention during these blank intervals. At no time were participants warned of a retest.

Each feedback trial lasted 9.5 sec and occurred in the following manner. First, the question–response–confidence

cue was presented for 4 sec. This consisted of a question responded to in the test phase, the participant's response to that question, and the participant's confidence in that response. Then, a central fixation crosshair was presented for 2.5 sec. Then, feedback was presented for 2 sec in the form of the correct answer in red (if the participant's response had been either incorrect or "xxx") or green (if the participant's response was correct). Finally, the screen then went blank for 1 sec between each feedback trial. The 4-sec blank intervals were also added at random between trials.

The first feedback trial presentation began 12 sec after the scanner began acquiring functional images to allow for tissue steady-state magnetization. A structural scan, which required 18 min, was acquired after all trials were complete.

The retest phase was administered out of the scanner and started approximately 5 min after the structural scan was complete. Participants were retested on the 140 questions for which they had received feedback. Participants gave responses (or entered "xxx") and rated their response confidence for each question. No feedback was given.

fMRI Recording

Images were acquired on a General Electric 1.5-T Twin Speed System with a standard GE head coil. A pillow and a tape were used to minimize head movement. Functional images were acquired in a single ~24-min run using an echo-planar sequence (repetition time = 3000 msec, echo time = 50 msec, flip angle = 90°). During this run, 478 sets of 32 contiguous 3.6-mm-thick axial images were acquired parallel to the anterior–posterior commissure plane (in-plane resolution = 3.125 × 3.125 mm). The first four volume acquisitions were dummy scans intended only to approach steady-state magnetization before the participant began the first trial in Stage 2. After the completion of the functional task and scan, structural images were acquired in the form of 124 contiguous 1.2-mm-thick axial images using an axial 3-D spoiled gradient-recalled sequence (repetition time = 3000 msec, echo time = 34 msec, flip angle = 45°, in-plane resolution = .9375 × .9375 mm). The structural scan took ~18 min to complete.

fMRI Analysis

Each participant's structural image was filtered with a SUSAN (smallest univalue segment assimilating nucleus) structure-preserving noise reduction filter (Smith & Brady, 1997). All subsequent preprocessing and analysis was accomplished with SPM99 (Wellcome Department of Imaging Neuroscience, London, United Kingdom). The functional images were corrected for slice acquisition timing, realigned, and coregistered with the structural images, and both the functional and structural images were then spatially normalized to the Montreal Neurological Institute (MNI) template.

Neural responses associated with the question–response–confidence cue and the feedback were modeled with rectangular functions (durations of 4 sec for the question–response–confidence cue and 2 sec for the feedback presentation) that were convolved with SPM99's canonical hemodynamic response function to create the design matrix for two different general linear models employed in the analysis.

The model, which will be referred to as the "Accuracy × Confidence" model, consisted of eight different event types: the question–response–confidence cue, the feedback to omit trials, and low/medium/high confidence errors/corrects (1 + 1 + 6 = 8 event types). Functional images were smoothed with a 6-mm FWHM kernel, and the model was estimated for each participant's data (49-sec high-pass filter).

The model was estimated for each participant's data as it was above, with the exception that each participant's contrast images were intensity normalized against the intercept image and then smoothed with a larger 10-mm FWHM kernel before entry into the group analysis. The effects of interest were (1) the contrast between the BOLD response to feedback for high confidence errors and low confidence errors, (2) the contrast between the response to feedback for low confidence corrects and high confidence corrects, (3) the contrast between 1 and 2, that is, contrasting activity elicited by metamemory mismatch for errors than by metamemory mismatch for corrects.

We also used a model designed to look for BOLD responses that distinguished between feedback to errors that were corrected at retest and those that were not corrected at retest. This model consisted of six event types: question–response–confidence cues; low confidence, medium confidence, and high confidence corrects; errors corrected at retest; and errors not corrected at retest. The effect of interest for this model was the contrast of errors corrected at retest with errors not corrected at retest.

Figures were generated with the program MRIcro (Rorden & Brett, 2000). All activations are superimposed over the average normalized structural images of all participants. All figures are displayed in neurological orientation. Figures show voxel activations reliable at $|t(13)| > 2$, for qualitative display of the results, and activations reliable at $|t(13)| >= 5.2$, for display of the statistical reliable results ($p < .05$, Bonferroni corrected for number of resolution elements, as estimated by SPM99). Finally, the Talairach Daemon anatomical labels and coordinates were determined using the MSU utility for SPM99, which was written by Sergey Pakhomov.

Behavioral Results

Each participant's mean response probability was weighed equally in the calculation of the across-participant mean probabilities displayed in Table 1.

Table 1. Conditional Probabilities (with *SEM*) of Responses

	p (Response Confidence)	*Given Actual Response Confidence*	*p (Correct at First Test)*	*Given Incorrect at First Test and Designated Response Confidence*	*p (Correct at Retest)*
Omit	.23 (0.03)	Omit	– (–)	Omit	.49 (0.05)
Low	.26 (0.04)	Low	.15 (0.02)	Low	.70 (0.05)
Medium	.25 (0.03)	Medium	.39 (0.03)	Medium	.78 (0.05)
High	.26 (0.04)	High	.78 (0.03)	High	.79 (0.05)

Basic Data

Participants correctly responded to an average of 34% of the questions at initial test. The first column of Table 1 shows the probability of each confidence level at first test and includes all trials, not only those selected for feedback presentation. The second column also includes all first test trials and shows that participants' confidence ratings were predictive of initial test accuracy: The mean gamma correlation between confidence and initial test accuracy was .64, $t(13) = 23.9$, $p < .001$ (for a description of the gamma correlation, see Nelson & Narens, 1980). Most of the 80 errors presented at retest were corrected; average retest accuracy was 69%.

Hypercorrection Effect

As in the previous experiments, participants were more likely to correct errors endorsed with higher confidence: The mean gamma correlation between confidence in an error of commission and retest accuracy was .14, $t(13) = 3.5$, $p < .005$.

fMRI Results

High Confidence Errors

The primary interest in this experiment was whether neural activity to the corrective feedback, as indexed by the BOLD response, would be modulated by the confidence with which the error was originally endorsed. This question was addressed by contrasting the activity elicited by feedback to high confidence error trials with that elicited by low confidence error trials (see Figure 1). Locations evidencing such differential activation included an area spanning the bilateral anterior cingulate gyrus and the right medial frontal gyrus. Another area modulated by error confidence spanned parts of the right middle frontal gyrus and right precentral gyrus or, broadly, the DLPFC. The third and final area was in the right inferior parietal lobule (IPL) or, broadly, the right TPJ. No areas evidenced significantly greater activity for low confidence errors than for high confidence corrects.

Low Confidence Corrects

Like errors, correct responses can reveal metamemory mismatch, as was shown both in Butterfield and Metcalfe's (2006) and in Butterfield and Mangels' (2003) experiments. Indeed, in those experiments, the responses were of about the same magnitude for error-related metamemory mismatch and for correct response metamemory mismatch. Metamemory mismatch elicited by feedback to correct trials is, of course, greater for low confidence correct responses than for high confidence correct responses. Our question was whether the same areas were activated for error-related and for correct-response metamemory mismatch. As is shown in Figure 2, the medial frontal gyrus was more active for low confidence correct responses than for high confidence correct responses. This area was similar, although somewhat more dorsal, to the area of ACC and medial frontal gyrus found in the error confidence analysis. No areas evidenced significantly greater activity for high confidence corrects than for low confidence corrects.

Omits

There was no significant difference in activations between the low confidence error responses and omit (or "xxx") responses. As might be expected from this result, when we contrasted omits to high confidence errors, the resultant pattern was highly similar to that found for the contrast between low confidence errors and high confidence errors. The same regions seen in latter contrast showed up at a lenient criterion in the former but did not reach the strict criterion for significance.

Differences in Activation for High Confidence Errors versus Low Confidence Corrects

This contrast addresses the question of whether there is additional processing of high confidence errors over and above that implied by metamemory mismatch per se. Contrasting metamemory mismatch for errors with metamemory mismatch for corrects revealed relatively more

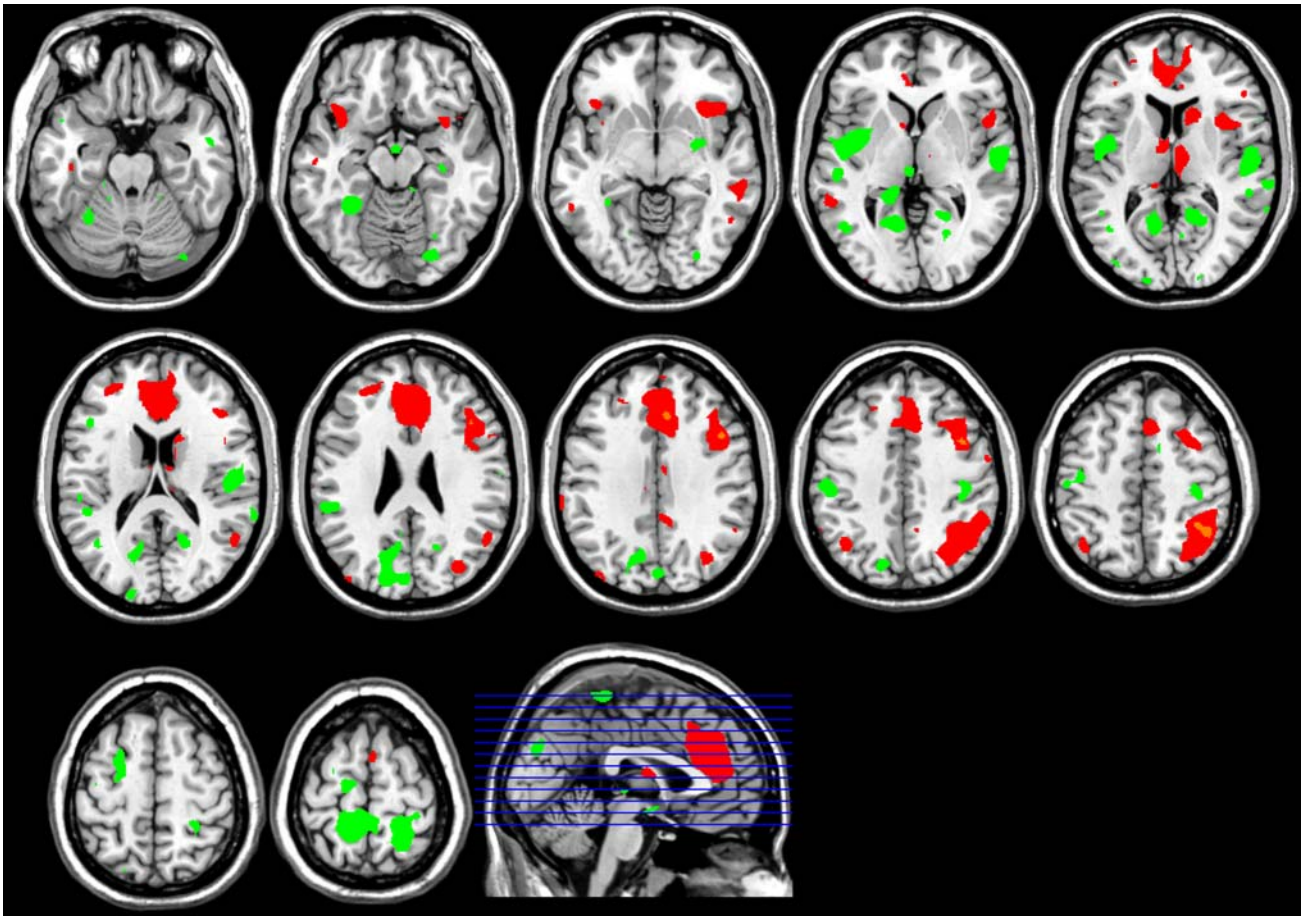

Figure 1. Contrast map of high confidence errors and low confidence errors. Red areas were more active for high confidence errors, whereas green areas were more active for low confidence errors; both thresholded at an uncorrected threshold of $p < .001$. Yellow and blue areas, respectively, denote areas of higher and lower activation for high confidence errors at a family-wise-corrected p level of .05. These more stringently thresholded areas are small and barely visible.

MNI Coordinates for x, y, and z of Cluster Peak (p < .05, Corrected)	***Population-based Incidence (%), Gray-matter Anatomical Label***	***Brodmann's Area***
6, 28, 32	89%, right anterior cingulate gyrus	32 (R)
44, 14, 36	49%, right precentral gyrus (location in white-matter for >50% of population)	9 (R)
46, −52, 50	67%, right IPL	40 (R)

Table of values corresponding to Figure 1.

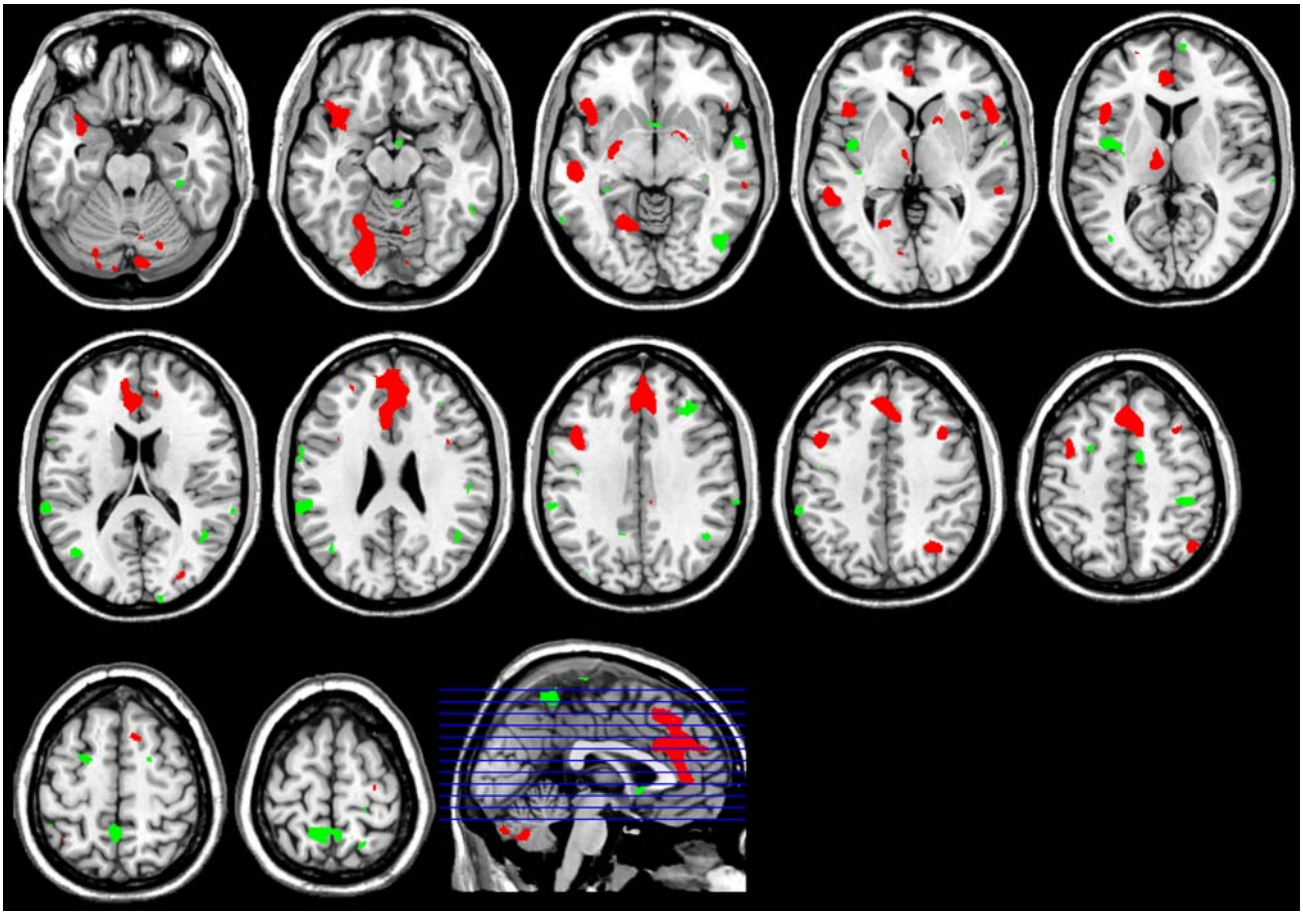

Figure 2. Contrast maps of low confidence corrects and high confidence corrects. Red areas were more active for low confidence corrects, whereas green areas were more active for high confidence corrects; both thresholded at an uncorrected threshold of $p < .001$. Yellow and blue areas, respectively, denote areas of higher and lower activation for low confidence corrects at a family-wise-corrected p level of .05. These more stringently thresholded areas are small and barely visible.

MNI Coordinates for x, y, and z of Cluster Peak (p < .05, Corrected)	Population-based Incidence (%), Gray-matter Anatomical Label	Brodmann's Areas
0, 34, 46	23%, midline superior frontal gyrus (location in white-matter for >50% of population)	8 (BL)
6, 48, 30	95%, right medial frontal gyrus	9 (R)

Table of values corresponding to Figure 2.

activity for error mismatch in the right DLPFC and the right IPL (TPJ), as is shown in Figure 3. These areas revealed more sensitivity to metamemory mismatch for errors than to metamemory mismatch for correct trials.

Error Correction

Finally, we looked for activation that distinguished between errors corrected at retest and errors not corrected

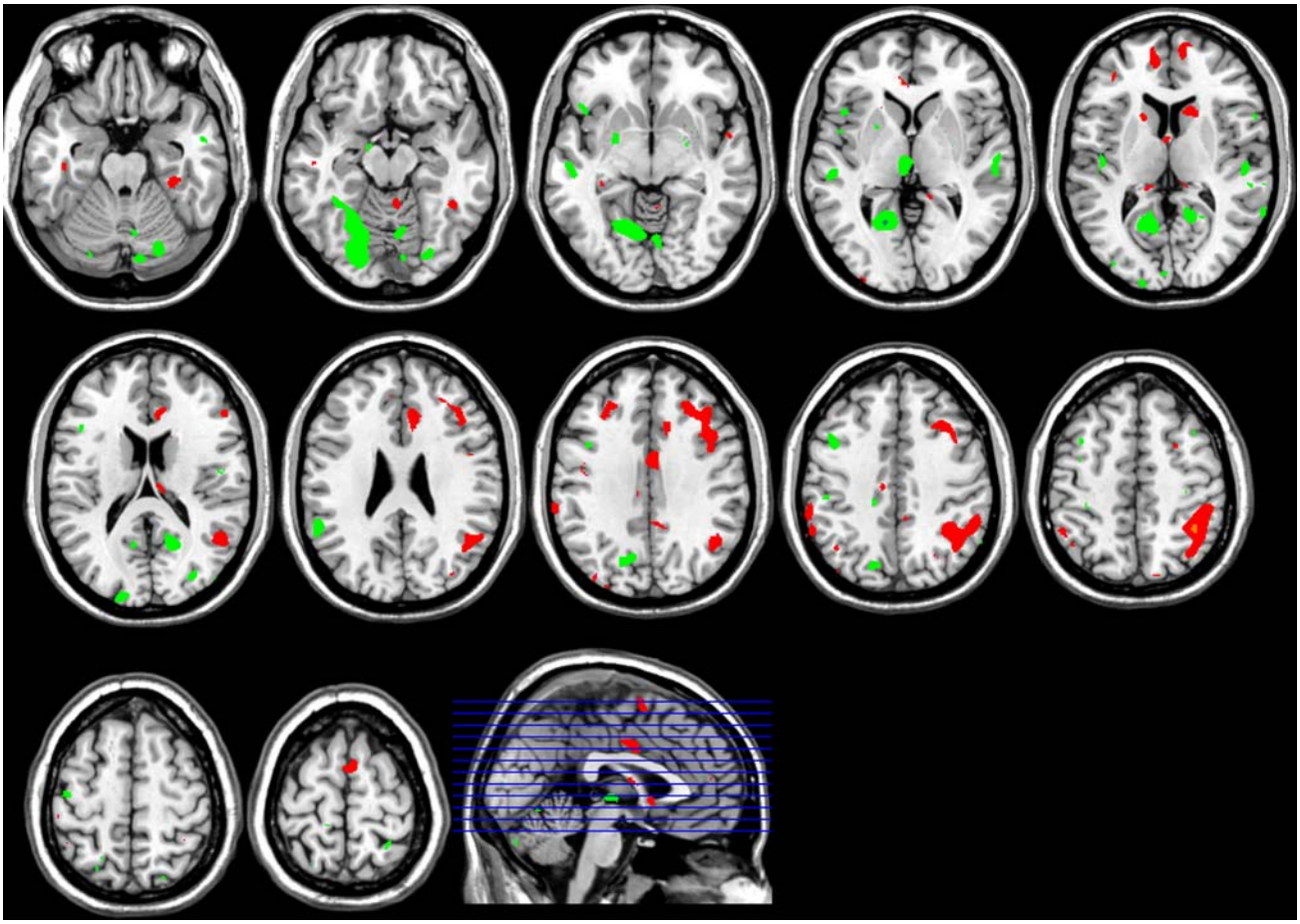

Figure 3. Maps of the contrast of high confidence errors and low confidence errors contrasted with the contrast of low confidence corrects and high confidence corrects. Red areas were more active for metamemory mismatch to errors than for metamemory mismatch to corrects, both thresholded at an uncorrected threshold of $p < .001$. Yellow and blue areas, respectively, denote areas of higher and lower activation for metamemory mismatch to errors at a family-wise-corrected p level of .05. These more stringently thresholded areas are small and barely visible.

MNI Coordinates for x, y, and z of Cluster Peak (p < .05, Corrected)	Population-based Incidence (%), Gray-matter Anatomical Label	Brodmann's Areas
28, 31, 34	53%, right middle frontal gyrus	9 (R)
40, −48, 50	IPL	40 (R)

Table of values corresponding to Figure 3.

at retest. No areas were significantly more active for items corrected at retest than for items forgotten at retest. The absence of significant results for this analysis may have occurred because of a lack of power. There were a number of areas associated with the reverse contrast—bilateral cingulate gyrus, paracentral lobule, precentral gyrus, postcentral gyrus, precuneus, medial frontal gyrus, insula, right middle frontal gyrus and putamen, and left superior temporal gyrus and middle temporal gyrus—perhaps revealing irrelevant and off-task processing.

DISCUSSION

A main cluster of brain activity was found to both the feedback to high confidence errors and the feedback to

low confidence corrects, which appears to relate to metamemory mismatch, as hypothesized. This area was a fronto-medial cluster spanning parts of bilateral ACC, bilateral medial frontal gyrus, and right cingulate cortex. The activation of this particular area is consistent with the idea that metacognitive mismatch is focally involved in the hypercorrection effect. Both metacognitive mismatch conditions showed activation in an area implicated in conflicting ideation. The activation observed in this area is also broadly consistent with the p3a deflection associated with metacognitive mismatch in the hypercorrection paradigm revealed in the earlier ERP (Butterfield & Mangels, 2003) investigation.

The medial frontal gyrus region, again, activated in both metacognitive mismatch conditions, has been implicated in the conscious monitoring of emotional states (Phan, Wager, Taylor, & Liberzon, 2002) including making personally relevant decisions or mental state attributions, especially attributions that are self-relevant (Jenkins & Mitchell, 2011; Powell, Macrae, Cloutier, Metcalfe, & Mitchell, 2009; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004), as well as in making metacognitive decisions (Fleming & Dolan, in press; Fleming, Weil, Nagy, Dolan, & Rees, 2010) including, but not limited to, self-relevant metacognitive decisions (Miele, Wager, Mitchell, & Metcalfe, 2011). That such processing should be implicated in error correction was not specifically expected in this paradigm insofar as error correction might be considered to be a cool (Metcalfe & Mischel, 1999) unemotional, non-self-involving cognitive process. However, participants in our experiments often reported being highly involved in their beliefs and sometimes reported some embarrassment, in the case of their high confidence errors, and delight, in the case of their low confidence correct responses. Thus, it is not inconsistent with the phenomenology sometimes reported in this task that people would be self-involved in the correctness of their stated beliefs. This connection to self-relevant emotional responding and the negative emotions that may be elicited upon hearing that one's high confidence responses were incorrect may also help explain the slight shift in location of activation in this cluster during feedback to high confidence errors as compared with low confidence corrects. The medial frontal activation to the mismatch on corrects was slightly more dorsal overall than the mismatch on errors. This difference may have occurred because people may have had a different emotional response to learning that they had been correct as compared with learning that they had been wrong. This distinction and its connection to variants in the cognitive nature as compared with the emotionality of the response has been underlined by Bush, Luu, and Posner (2000); Steele and Lawrie (2004); and Amodio and Frith (2006), who have provided detailed descriptions and a meta-analyses of these shifts in function.

As well as the common frontal areas, which confirm the metacognitive mismatch hypothesis, there were two additional areas that were sensitive only to trials on which errors were committed. The dorsolateral prefrontal activation was associated selectively with the processing of high confidence error feedback and not with low confidence correct feedback. One primary difference between the two metacognitive mismatch conditions is, of course, the presence of the wrong answer, in the error correction case. To prepare for future correct responding, the person who had made an error not only has to encode the correct answer more fully (as in the correct answer case) but also has to discriminate the correct answer from the wrong answer and suppress the wrong answer. Neither discrimination nor suppression is needed in the low confidence correct response case.

The right DLPFC, which is associated with both suppression and response discrimination, was selectively activated in the error condition but not the correct responding condition. MacDonald et al. (2000), in a highly influential model, have also proposed that the DLPFC acts as a controller that resolves conflict. There was a conflict of representations only in the error condition, of course. In the case of the hypercorrection situation, presumably, the conflict would be between the incorrectly retrieved answer and the presented correct answer. Hayama and Rugg (2009) have argued that the right DLPFC plays a role in post-retrieval monitoring more generally. Both the correct and error responses would have been in a postretrieval state. Certainly, some kind of monitoring and discrimination between the correct and incorrect responses is needed, and this role for the DLPFC is also consistent with the present paradigm. And, as noted earlier, Anderson and his colleagues (Anderson & Levy, 2006; Anderson et al., 2004) have investigated the specific role of the DLPFC in the suppression of unwanted memories—a possibility that seems eminently plausible in the present case, but only when an error has been committed. According to their research, when people do not want to think about certain memories, they can inhibit their retrieval—a process that their results show—implicates the right DLPFC. Suppression results in impaired retention of those unwanted memories. It is plausible that the DLPFC activation we observed in our task, selectively in the error correction condition, reveals the engagement of this suppression mechanism to undermine memory for the originally retrieved errors. Their suppression, as well as the enhanced attention to the correct response, would allow mnemonic processing to focus selectively on the correct answers and reveal a hypercorrection pattern in the data.

The behavioral data on whether suppression (which involves forgetting of the error) or mediation (which could involve strategic remembering of the error as a cue to getting to the target), either of which would involve DLPFC activation, was at play in hypercorrection are, unfortunately, inconclusive to date. Butterfield and Mangels (2003) and Butterfield and Metcalfe (2001) both asked participants to produce three responses on the final test and put a star beside the response that was correct. If suppression of the original error was necessary for error correction,

then a negative contingency between the production of the original error and correct responding should have emerged. If a mediation process by which the person used the error to help retrieve the target had occurred, then there should have been a positive contingency between the production of the original error and correct responding. Unfortunately, neither the suppression nor the mediation hypothesis was supported, insofar as it made no difference for correct responding, whether the original error was mentally present or not at time of test. So, we are not in a position to infer whether the DLPFC activation was because of discrimination or suppression. Even so, both suppression and discrimination (needed for mediation to be effective) are implicated in this region, and perhaps, both were at work. (It is conceivable that more detailed event-specific future imaging could pinpoint the locale of these two different processes.)

Finally, the right TPJ—taken broadly, as in the work of Van Overwalle (2009), Decety and Lamm (2007), and others, to include the IPL—was selectively activated to the feedback to high confidence errors. This is an area that is activated in TOM tasks (Zaitchik et al., 2010; Van Overwalle, 2009; Decety & Lamm, 2007; Saxe & Powell, 2006; Saxe & Wexler, 2005; Saxe & Kanwisher, 2003; although, note that the peak activation in the present hypercorrection paradigm was about 25 mm dorsal to the peak of the story-based TOM activation identified in a recent meta-analysis by Mar, 2011). Although this general TPJ region is also implicated in other tasks such as empathy (e.g., Jackson, Brunet, Meltzoff, & Decety, 2006), reorienting (e.g., Corbetta & Shulman, 2002), and agency detection (e.g., Miele et al., 2011; Farrer & Frith, 2002), its involvement in the error correction task seems most akin to that in the TOM tasks.

Like the error correction task investigated here, classical TOM tasks involve the consideration of alternative beliefs—those of the self compared with those of the other. In the error correction paradigm, it is one's own belief that is false, and one has to accept and remember that the response given by the computer is the correct answer. In a typical TOM task, one's own beliefs are considered to be true, and those of the other, false. For example, one may have experienced that there are pencils in a Smartie's box but needs to realize that another person would think that it contains candies. In the hypercorrection paradigm, it is also a computer holding the alternative "belief," although it seems likely that most people attribute the answers given by the computer to the experimenter, who is a person. Aside from who is right and wrong, then, the structure of the false belief task and the error correction task seem to be fairly similar.

Many researchers have interpreted ToM tasks as implicating more than coordination of two sets of information, however. Additionally, attributions having to do with an understanding that other people have minds and that their minds can hold different beliefs from one's own mind are often evoked (Decety & Sommerville, 2003; Ruby & Decety, 2001). It is not clear whether the presence of any other mind is implicated in the error correction paradigm. If it were, presumably, it would be a "mind" reflecting what people in general believe (or perhaps the mind of the experimenter). Thus, although the informational content of the TOM tasks and the error correction tasks seem to be closely related, the central mentalizing aspects of the two tasks, which provide much of the interest in ToM studies, may be different. Alternatively, although, and speculatively, it is possible that the process of correcting one's own erroneous beliefs is a social process that involves understanding that one's own beliefs can be wrong and that anothers' can be right and that updating one's own erroneous beliefs relies on adopting and accepting the perspective of others.

Acknowledgments

We thank Joy Hirsch and Steve Dashnaw. This research was supported by NIMH grant F31MH66663 and NIA grant R01AG26158.

Reprint requests should be sent to Janet Metcalfe, Department of Psychology, Columbia University, 401B Schermerhorn Hall, New York, NY 100127, or via e-mail: jm348@columbia.edu.

Note

1. For example, if a participant had answered all 599 questions and had only given 15 low-confidence corrects (which are, for most participants, the rarest trial type), the program "grew" the low-confidence category from "below −17" to "below −16" and so on until there were 20 low-confidence corrects. If there were now a shortage of medium-confidence errors or corrects, this category was similarly grown until there were at least 20 trials of medium-confidence corrects and medium-confidence errors.

REFERENCES

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277.

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410,* 366–369.

Anderson, M. C., & Levy, B. J. (2006). Encouraging the nascent cognitive neuroscience of repression. *Behavioral and Brain Sciences, 29,* 511–513.

Anderson, M. C., Ochsner, K., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., et al. (2004). Neural systems underlying the suppression of unwanted memories. *Science, 303,* 232–235.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology, 58,* 97–105.

Bunge, S. A., Hazeltine, E., Scanlon, M. D., Rosen, A. C., & Gabrieli, J. D. E. (2002). Dissociable contributions of prefrontal and parietal cortices to response selection. *Neuroimage, 17,* 1562–1571.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Science, 4,* 215–222.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2010). *The hypercorrection effect persists over a week, but high confidence errors return*. Poster presented at the 2010 Psychonomics Society Annual Meeting, #1102.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918–928.

Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research, 17,* 793–817.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491–1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1,* 1556–1623.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus driven attention in the brain. *Nature Neuroscience Reviews, 3,* 201–215.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist, 13,* 580–593.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Science, 7,* 527–533.

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs. another person as being the cause of an action: The neural correlates of the experience of agency. *Neuroimage, 15,* 596–603.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin and Review, 16,* 88–92.

Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological Science, 21,* 801–803.

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition, 38,* 951–961.

Fleming, S. M., & Dolan, R. J. (in press). The neural basis of accurate metacognition. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329,* 1541–1543.

Hayama, H., & Rugg, M. D. (2009). Right dorsolateral prefrontal cortex is engaged during post-retrieval processing of both episodic and semantic information. *Neuropsychologia, 47,* 2409–2416.

Huelser, B. J., & Metcalfe, J. (2011). Masking related errors facilitates learning, but learners do not know it. *Memory & Cognition*.

Hunt, R. R. (2009). Does salience facilitate longer term retention? *Memory, 17,* 49–53.

Jackson, P. L., Brunet, E., Meltzoff, A. N., & Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain: An event-related fMRI study. *Neuropsychologia, 44,* 752–761.

Jenkins, A. C., & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience, 6,* 211–218.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

Koriat, A. (1997). Monitoring one's knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition, 36,* 416–428.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51,* 481–537.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279–308.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68,* 522–528.

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General, 118,* 100–104.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288,* 1835.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. A. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14,* 647–654.

Mar, R. A. (2011). The neural bases o f social cognition and story comprehension. *Annual Review of Psychology, 62,* 103–134.

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1084–1097.

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high confidence errors: Did they know it all along? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 437–444.

Metcalfe, J., & Finn, B. (in press). Hypercorrection of high confidence errors in children. *Cognition and Instruction*.

Metcalfe, J., & Mischel, W. (1999). A hot/cool system analysis of delay of gratification: Dynamics of willpower. *Psychological Review, 106,* 3–19.

Miele, D. B., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience, 23,* 3620–3636.

Murdock, B. B., Jr. (1974). *Human memory: Theory and data*. Potomac, MD: Erlbaum.

Nee, D. E., & Jonides, J. (2008). Dissociable interference control processes in perception and memory. *Psychological Science, 19,* 490–500.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning & Verbal Behavior, 19,* 338–368.

Paller, K. A., McCarthy, G., & Wood, C. C. (1988). ERPs predictive of subsequent recall and recognition performance. *Biological Psychology, 26,* 269–276.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8.

Phan, K. L., Wager, T. D., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16,* 331–348.

Picton, T. W. (1992). The P300 wave of the human event-related potential. *Journal of Clinical Neurophysiology, 9,* 456–479.

Powell, L. J., Macrae, C. N., Cloutier, J., Metcalfe, J., & Mitchell, J. P. (2009). Dissociable neural substrates for agentic versus conceptual representations of self. *Journal of Cognitive Neuroscience, 22,* 2186–2197.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 803–814.
Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioral Neurology, 12,* 191–200.
Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience, 4,* 546–550.
Rumelhart, D. E., & McClelland, J. L. (1987). *Parallel distributed processing*. Cambridge, MA: MIT Press.
Saxe, R. (in press). The right temporo-parietal junction: A specific brain region for thinking about thoughts. In A. Leslie & T. German (Eds.), *Handbook of theory of mind*. Psychology Press.
Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." *Neuroimage, 19,* 1835–1842.
Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17,* 692–699.
Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43,* 1391–1399.
Sitzman, D. M., & Rhodes, M. G. (2010). *Does the hypercorrection effect occur when feedback is delayed?* Poster presented at the 2010 Psychonomics Society Annual Meeting, #5090.
Smith, S. M., & Brady, J. M. (1997). SUSAN—A new approach to low level image processing. *International Journal of Computer Vision, 23,* 45–78.
Steele, J. D., & Lawrie, S. M. (2004). Segregation of cognitive and emotional function in the prefrontal cortex: A stereotactic meta-analysis. *Neuroimage, 21,* 868–875.
Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30,* 829–958.
Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., & Dickerson, B. C. (2010). Mental state attribution and the temporoparietal junction: An fMRI study comparing belief, emotion, and perception. *Neuropsychologia, 48,* 2528–2536.