Metacognition Learning (2006) 1: 69–84
DOI 10.1007/s11409-006-6894-z

EMPIRICAL STUDY

The correction of errors committed with high confidence

Brady Butterfield · Janet Metcalfe

Received: 9 December 2005 / Revised: 14 December 2005 /
Accepted: 19 December 2005 / Published online: 23 March 2006

Abstract Most theories predict that when people indicate that they are highly confident they are producing their strongest responses. Hence, if such a high confidence response is in error it should be overwritten only with great difficulty. In contrast to this prediction, we have found that people easily correct erroneous responses to general information questions endorsed as correct with high-confidence, so long as the correct answer is given as feedback. Three potential explanations for this unexpected hypercorrection effect are summarized. The explanation that is tested here, in two experiments, is that after a person commits a high-confidence error the correct answer feedback, being surprising or unexpected, is given more attention than is accorded to the feedback to low-confidence errors. This enhanced attentional capture leads to better memory. In both experiments, a tone detection task was presented concurrently with the corrective feedback to assess the attentional capture of feedback stimuli. In both, tone detection was selectively impaired during the feedback to high confidence errors. It was also negatively related to final performance, indicating that the attention not devoted to the tone detection was effectively engaged by the corrective feedback. These data support the attentional explanation of the high-confidence hypercorrection effect.

Keywords Metamemory · Metacognition · Confidence judgments · Error correction · Hyper-correction · Tone detection · Attention · Feedback

Errors are common in all realms of human cognition. Within the domain of memory, errors include failures of information retrieval, as well as the erroneous retrieval and endorsement of false information. Given the pervasiveness of errors in all realms of human cognition it is surprising how inadequate is our understanding of how we are able to avoid (c.f., Dodson, Koutstaal, & Schacter, 2000; Gallo, McDermott, Percer, & Roediger, 2001), and, especially, to overcome them. Nevertheless, the question of how people learn hinges critically on how they are able to replace misinformation with correct information, or, to put it more simply, how they are able to correct their errors. We review the literature and conduct two

B. Butterfield · J. Metcalfe (✉)
Department of Psychology, Columbia University, New York, NY 10027, USA
e-mail: metcalfe@psych.columbia.edu

new experiments that investigate the cognitive basis underlying how errors in retrieval from semantic memory, once made, can be subsequently corrected. In particular, the experiments focus on the relationship between the metamemory one has concerning the predicted accuracy of the answer, as reflected by response confidence, and the likelihood that the corrective feedback will be successful.

There is a long tradition in which it is considered that the best learning strategy is the complete avoidance of errors. There is also a large literature on how incorrect and/or misleading information worsens memory (Ayers & Reder, 1998; Loftus, 1979; Wright & Loftus, 1998). If self-generated errors are like experimenter-presented misinformation, then predictions can be made about how different kinds of errors—for example strong, highly confident errors as opposed to, say, errors about which the participant voices considerable uncertainty—will differentially impact later performance.

There is a considerable research literature, going back to Müller and Schumann (1894), Webb (1917), Melton and Irwin (1940), McGeoch (1942), Osgood (1949) and Barnes and Underwood (l959) through Loftus (1979) and J. R. Anderson and Reder (1999), showing that competing information results in interference. There are a variety of theories about why and how competing information produces decrements in memory (e.g., Atkinson & Shiffrin, 1968; J. A. Anderson, 1973; J. R. Anderson & Bower, 1972; Gillund & Shiffrin, 1984; Hintzman, 1984; Metcalfe, 1990; Metcalfe Eich, 1982; Raaijmakers & Shiffrin, 1981), but there is little doubt that such interference phenomena exist. Thus, when a person makes a mistake by generating a high-confidence error, that self-generated misinformation should harm the remembrance of the correct answer when it is subsequently presented.

Finally, there is general agreement that, except under very circumscribed circumstances, responses about which a person is highly confident tend to be items that are the strongest and most fluent in memory, though confidence does not appear to be scaled directly from perceived familiarity (see Van Zandt, 2000). 'Strength' is used here as a shorthand, without ascribing to either a unidimensional view of items or a particular scaling assumption. High confidence items are, presumably, the most dominant and easily retrievable items—even if objectively they are mistakes. Items about which the person exhibits low confidence are presumably less strong. Indeed, there is good evidence that retrieval fluency is a causative factor in confidence ratings (e.g., Kelley & Lindsay, 1993) and in other memory judgments (e.g., Jacoby, Woloshyn, & Kelley, 1989). By this reasoning, it should be most difficult to overcome highly confident errors, because the reasons for the high confidence in the errors are presumably the very factors that make those errors most likely to interfere with the learning of the correct response.

To evaluate the hypothesis that high confidence errors are less likely to be corrected than errors endorsed with less confidence Butterfield and Metcalfe (2001a) had participants answer general information questions. After each response the participant was asked to rate his or her confidence that the response was correct. Then, if the answer was incorrect, the participant was shown the correct answer. At a later time, participants were retested on some of the questions that elicited errors as well as some that elicited correct answers.

The core question concerned the relationship between initial confidence in an error and the likelihood of answering the same question correctly at retest. The first hypothesis was that errors endorsed with higher confidence would be less likely to

be corrected at retest than would lower confidence errors. As it was also of interest whether these high confidence errors were 'stronger,' participants gave three retest responses, wrote them down, and put a star beside the correct one. The second hypothesis was that higher confidence errors, as stronger errors, would be more likely to appear in this list of three responses than should lower confident errors. Thus, this experiment was concerned not only in whether a participant *could* think of a response (the correct one, say) but also in whether a participant *could not help but* think of a response (the original error, in particular).

As shown in Table 1 there was a systematic relation between confidence and accuracy at initial test, with higher confidence being related to high accuracy. This result suggests that confidence was an indication of something that might be ascribed the shorthand term 'memory strength' (c.f. Van Zandt, 2000).

Table 1 also shows the relation between confidence in an error and the likelihood of error correction at retest. In contrast to our major hypothesis, high confidence in the initial error predicted that the *correct* response would be both produced and subscribed to (starred) on the retest. Questions on which people had made high confidence errors were more (not less) likely to be answered correctly at retest than were low-confidence errors. The tendency to correct highly confident mistakes was not due to the absence of those mistakes from memory. The third column of Table 1 shows that higher confidence errors were more likely than lower confidence errors to show up at retest.

Explanations of the Hypercorrection Effect

Although the high-confidence hypercorrection effect was unexpected, a priori, as with many such counterintuitive findings, we were able to quickly generate possible explanations once the data were in. Here, we review some previously published data and also present new data bearing on three such explanations: the mediation explanation, the semantic neighborhood/familiarity explanation, and the enhanced attention explanation. We then present the results of two new experiments testing the third of these possible explanations.

Table 1 Conditional probabilities [with *SEM*] of responses from Butterfield and Metcalfe (2001b)

P(response confidence)	Given resp. conf.:	P(correct at retest)	Given incorrect at 1st test and resp. conf.:	P(correct at retest)	P(error in retest response set)	Mean normative ease
Low 0.52 [0.02]	*Low*	0.06* [0.01]	*Low*	0.61** [0.03]	0.51** [0.05]	0.19** [0.02]
Med. 0.08 [0.01]	*Med.*	0.39* [0.04]	*Med.*	0.81** [0.07]	0.86** [0.07]	0.18** [0.03]
High 0.41 [0.01]	*High*	0.89* [0.01]	*High*	0.87** [0.07]	0.77** [0.08]	0.28** [0.04]

*mean from the 71 participants with data in all three cells.

**mean from the 23 participants with data in all three cells.

High confidence errors might act as mediators to the correct response. To investigate the contingencies between correct performance and the appearance of the error at test we reanalyzed the data from Butterfield and Metcalfe (2001b). Confidence in an error might be positively correlated with both correct performance at retest and the presence of the error at retest. This pattern of results would suggest a relation between the presence of the correct answer and the original error at retest and provide support for the idea that participants were remembering correct answers by associating them with their original errors (i.e., mediating).

To evaluate this possibility we calculating a performance-adjusted difference score for each participant—the difference between performance with the error present in the response list and performance with the error absent, divided by total performance.[1] The average contingency score was .14, which means that a participant was on average 14% more likely to correctly answer a question at retest when the error given to that question initially was in the list of responses than when it was not. This value was reliably different from zero, $t(72) = 2.4$, $p < .05$. Thus, the presence of the participant's original error in the set of three responses reliably, if weakly, predicted the presence of the correct answer in the same set of responses.

Could mediation be a sufficient explanation of the hypercorrection effect? If so, one would expect hypercorrection to be minimal in analyses that held the presence of the error in the retest response list constant. The 29 participants who had suitable data (variance in retest accuracy and initial error confidence for both trials in which the error was absent from the retest response set and trials in which the error was present in the retest response set), still demonstrated overall hypercorrection, $\gamma = .33$, $t(28) = 3.4$, $p < .005$. Significant hypercorrection persisted even when the analysis was limited to trials in which the error was absent from the retest response set, $\gamma = .41$, $t(28) = 3.2$, $p < .005$, and when limited to trials in which the error was present in the retest response set, $\gamma = .37$, $t(28) = 2.3$, $p < .05$. Thus, the hypercorrection persisted even when the trials were limited to those in which mediation may have been happening and trials in which it definitely was not. If mediation is playing a role in the hypercorrection effect it is likely a small one, as the effect did not weaken in analyses that held it constant (both $t < 1$).

The semantic landscape, or the a priori familiarity, might differ for high confidence errors. Confidence ratings may be influenced by how much domain-relevant information a question activates (see, e.g., Glenberg, Wilkinson & Epstein, 1982) 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 an error concerning the capital of Canada), might increase the likelihood that the erroneous response is endorsed with high

[1] Evaluating the degree to which participants mediate is complicated by the possible non-independence of the two events, as the presence of a certain response leaves only two response slots for the other response type. This issue may be less relevant in this case, however, because on most trials participants had trouble coming up with a third response at all.

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.

We reanalyzed the data from Butterfield and Metcalfe (2001b) to investigate this possibility. The Nelson and Narens (1980) questions we had used as stimuli were normed. We used this normative item difficulty as a proxy for domain familiarity (easier questions, presumably, probe areas of higher average domain familiarity). The normative ease of a question was indeed correlated with the confidence of errors that question elicited, $r = .21$, $t(71) = 5.67$, $p < .001$ (see the far right column of Table 1, higher value = easier question). The normative ease of a question answered incorrectly at initial test was also correlated with the likelihood of that question being correctly answered at immediate retest, $r = .19$, $t(68) = 6.08$, $p < .001$). To see if confidence was related to error correction beyond what could be explained by normative difficulty, correlations that partialed out normative difficulty were calculated for each participant. The mean partial correlation that eliminated item difficulty was small, but reliably different from zero, $pr = .11$, $t(68) = 3.70$, $p < .001$. Thus, although normative difficulty is a substantial and significant covariate, it does not sufficiently explain the hypercorrection of high-confidence errors.

Because the questions presumably differ on dimensions other than normative difficulty, a true item-analysis was also calculated. Within-item (across-participant) gamma correlations between error confidence and retest accuracy were computed for questions that varied in the confidence of errors they elicited and their correction at retest. This item analysis revealed a significant correlation between confidence and correctness at initial retest, $\gamma = .22$, $t(63) = 2.08$, $p < .05$. Thus, item effects do not appear to be sufficient cause of the hypercorrection of high-confidence errors.

The results of our (2001b) study suggested that high-confidence errors were indeed 'stronger' than were low-confidence errors, as they were more likely to show up as one of the three retest responses. Though the primary intuition was that this 'strength' would interfere with the learning of the correct answer, these high-confidence errors were actually more likely to be corrected than were low-confidence errors, and this did not seem to be due to participants' associating the correct answer with these strong errors.

Though difficulty does not directly map to an individual's domain familiarity (as people know different topics to differing degrees, and there exist difficult questions about easy domains and vice versa), the correspondence is reasonable. To the extent that this equating of domain familiarity and difficulty is valid, the significant correlation between the normative ease of a question and the confidence of errors given to a question suggests that more domain familiarity may indeed correspond with higher confidence errors.

Butterfield and Mangels (2003) assessed the role of familiarity more directly by prompting subjects to rate the correct answer feedback as familiar or unfamiliar. Questions eliciting high-confidence errors were more likely to have answers

familiar to participants than were questions eliciting low-confidence errors, which were more likely to have answers unfamiliar to the participant. Familiar feedback was also more likely to be successfully generated at retest than was unfamiliar feedback. Thus, familiarity was contributing to the hypercorrection effect.

Enhanced attentional capture following a high confidence error. One final process that might explain the hypercorrection effect is that feedback indicating that one's highly confident response was incorrect might well be more surprising to an individual than feedback indicating that an incorrect response given with less confidence was incorrect. Many experiments have shown that novel or surprising events are better remembered than unsurprising events (Hunt & Lamb, 2001; von Restorff, 1933; Sokolov 1963; Tulving & Kroll, 1995), presumably because they capture more attention than do their more mundane counterparts. A novelty-related boost in encoding of the correct response would be expected to enhance memory. This enhancement may suffice to counteract the purported a priori associative strength of highly confident incorrect responses. The monitoring/feedback loop in the CHARM model is one conceptualization of a process that would create this memory enhancement (see Metcalfe, 1993a).

Although there is as yet no direct evidence that there is attentional enhancement when people are given feedback to high-confidence errors, the event-related potential experiments conducted by Butterfield and Mangels (2003) lend such an explanation plausibility. Analysis of a fronto-central positivity peaking approximately 350 milliseconds after accuracy feedback presentation revealed a significant effect of error confidence, such that it was larger for high-confidence errors than for low-confidence errors. The amplitude of this component also reliably predicted error correction at subsequent retest: it was larger for items later generated at retest than items later missed.

This component is analogous to the novelty-P3/P3a, a frontally-maximal waveform elicited by rare target stimuli and other events that are novel in a given context (e.g., Courchesne, Hillyard, & Galambos, 1975; Goldstein, Spencer, & Donchin, 2002; Knight, 1984). Given that high-confidence errors are not only rare, but also presumably more arousing than other types of errors, it seems plausible that the frontal component might have been signaling increased attention to the feedback to high-confidence errors, relative to low-confidence errors, resulting in enhanced memory (see Metcalfe, 1993b).

The current experiments investigate these relationships more directly by including a secondary tone-detection task along with the general-information question task. We expected the tone detection task to elicit differential performance as a function of the amount of attention that was engaged elsewhere. It has been found, for example, that pain disrupts tone detection performance (Crombez, Eccleston, Baeyens, & Eelen, 1996), and that distracting visual stimuli raise auditory thresholds. It has also been found that a stimulus can impose a cross-modal "attentional-blink" that impairs the encoding of a subsequent stimulus of a different modality (Jolicoeur, 1999). A number of dual-task studies have found that tone-detection is impaired when performing another task (e.g., Madden, 1986 and see Arnell, 2001, for a review of cross-modal attentional capture).

In the current experiments, participants engaged in a general-information task similar to that employed previously. To assess the amount of attention captured by the feedback presentation, feedback presentation was often accompanied by a to-be-detected tone. It was hypothesized that feedback to high-confidence errors and low-confidence corrects (conditions of expectation violation) would be associated with impaired tone detection performance relative to feedback to low-confidence errors and high-confidence corrects (conditions of expectation confirmation). The enhanced attentional capture associated with such feedback would, presumably, leave less attention to be devoted to the tone presentation. It was further posited that if the hyper-correction effect were attributable to attentional differences, then there would be a negative relation between tone detection and performance on the final test (such that people should do better on a question when they failed to detect the tone) because that meant that their attention was instead engaged on the feedback.

Experiment 1

Method

Participants

Thirty-nine Columbia undergraduates (23 women and 16 men, mean age 21 years) participated to partially satisfy an Introductory Psychology course requirement. The data from three participants who neglected the tone detection task were excluded (participants detected fewer than 8% of the feedback tones overall). Participants were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992).

Materials

619 general information questions taken from sources including the set published by Nelson and Narens (1980), various board games, and internet trivia sites were used as stimuli. A sample question was "What igneous rock makes up the bulk of Devils Tower?" (answer: "basalt"). All correct answers were a single word.

Procedure

The experiment consisted of two phases, an initial test and a surprise retest. The initial test included a tone detection task, but the retest did not. In both tests, general information questions were presented in the center of the computer screen and the participant was given an unlimited amount of time to type a response. If participants were not certain about the answer, they were encouraged to make an educated guess. If they felt that they could not come up with a remotely plausible answer, they were instructed to type "xxx." After each

response, participants rated their confidence in the accuracy of their response on an analog horizontal sliding scale ranging from sure incorrect (left), to unsure (middle), to sure correct (right). The slider bar was initially at the middle, unsure point. Participants used a mouse to click on this slider bar and drag it to the left or right to make their confidence rating. The program coded each confidence rating as an integer between −50 (sure wrong) and +50 (sure right). In the instructions, participants were encouraged to use the entire scale. Though participants rated their confidence in omit ("xxx") responses to keep trials consistent, these ratings were automatically considered to be −50.

Immediately following the confidence rating, feedback was presented for 1.5 seconds: the correct answer presented in green if it matched the participants' response, or in red if it did not. The program used a letter-matching algorithm to score the response between 0 and 1. Each response was scored as correct (.75 or greater), incorrect (less than .70), or borderline (greater than or equal to .70 and less than .75—pilot work determined that these responses were either incorrect or badly misspelled correct responses). Feedback to "xxx" 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.

The initial test phase consisted of this general information task and a simultaneous tone detection task. Before beginning the experiment, the volume was adjusted to a level at which the participant reported s/he could, with some difficulty, hear the tone. In the instructions for the task, participants were asked to hit the spacebar every time they heard the tone. The tone was 250 Hz and 100 milliseconds in duration. For each trial, there was a 25% chance that the tone was presented starting at a random time between 100 and 1900 milliseconds after the question was presented. These tones were presented to introduce an element of uncertainty about when the to-be-detected tones would occur, but are not otherwise of interest.

The tones presented during feedback were the tones of critical interest to the experimental question. These were presented at a random time between 100 and 500 milliseconds after feedback presentation onset. The probability of a feedback tone being presented was a function of response accuracy and response confidence. Participants' confidence ratings are generally accurate, so high-confidence errors and low-confidence corrects occur infrequently. Medium-confidence responses are also relatively infrequent. It was desirable for the tones to be presented infrequently enough that participants did not come to expect the tone on every trial, but a sufficient number of tone presentations in the sparse trial conditions (high-confidence errors and low-confidence corrects) was required to provide enough data to test the hypotheses. Therefore, the scale of confidence, which ranged from −50 to +50, was split into low confidence (confidence less than −17), medium confidence (confidence between −17 and +17, inclusive) and high confidence (confidence greater than +17) categories. The probability of tone presentation, during feedback, was assigned as follows: For the most frequent responses—the low-confidence errors, the high-confidence corrects, and the omits—the probability of tone presentation was .25. For the responses of medium frequency of occurrence—the medium-confidence corrects and the medium

frequency incorrects—the probability of tone presentation was .75. For the responses of low frequency—the high-confidence errors and the low-confidence correct responses—the probability of tone presentation was 1.0.

The initial test began with five practice trials (during which on-screen instructions appeared) and lasted a total of 25 minutes. After the initial test and a 2-minute interpolated task, a surprise retest consisting of questions answered incorrectly in the initial test was given. Retest trials were identical to the initial test trials, except that there was no tone detection task.

Results

Each participant's response probability was weighted equally in the calculation of the mean probabilities displayed in Table 2. Though this table only includes data from participants who had trials in all cells in a given column, the within-participant gamma correlations were computed for all participants with variance in the two measures being correlated.

Basic Data

The general information questions were difficult, and participants responded to an average of 25% of them correctly at initial test. Participants gave omit responses on an average of 38% of the trials. These trials will not be included in any analyses. Among non-omit responses, participants responded to an average of 41% of questions correctly at first test. Most errors of commission were corrected at retest—average retest accuracy for non-omits was 77%.

Participants' confidence ratings were predictive of initial test accuracy: the mean gamma correlation between confidence and initial test accuracy was .63, $t(35) = 29.2$, $p < .001$.

Tone detection performance was better when the tone was presented during the question than when it was presented during the feedback. On trials in which participants did not give an omit response, they detected 92% of the question tones and 73% of the feedback tones, and this difference was significant, $t(35) = 6.5$, $p < .001$. Any mention of "tones" in the remainder of the results section refers to the tones presented during feedback (the tones relevant to the experimental question).

Table 2 Conditional probabilities [with *SEM*] of responses, Experiment 1

P(response confidence)	Given resp. conf.:	P(correct at 1st test)	Given incorrect at 1st test and resp. conf.:	P(correct at retest)
Omit 0.38 [0.03]	*Omit*	- [-]	*Omit*	0.48** [0.03]
Low 0.23 [0.02]	*Low*	0.16* [0.02]	*Low*	0.73** [0.03]
Med. 0.15 [0.02]	*Med.*	0.34* [0.03]	*Med.*	0.76** [0.04]
High 0.25 [0.02]	*High*	0.70* [0.03]	*High*	0.82** [0.03]

*mean from the 34 participants with data in all four cells.

**mean from the 33 participants with data in all four cells

The Hyper-correction Effect

This experiment replicated the hypercorrection effect: errors endorsed with high confidence were more likely to be corrected at retest than errors endorsed with low confidence. The mean gamma correlation between confidence in the original errors of commission and retest accuracy was significantly positive, $\gamma = .13$, $t(35) = 2.5$, $p < .05$.

Tone Detection and Confidence

As hypothesized, increased metamemory mismatch was associated with impaired tone detection (see Fig. 1). For errors, the mean gamma correlation between error confidence and tone detection for trials on which a tone was presented was $-.22, t(35) = -3.7$, $p < .001$. This result is consistent with the notion that feedback to a high-confidence error captured more attention than did feedback to a low-confidence error, and that this attention capture was accomplished at the expense of performance on the secondary task.

For correct responses, the correlation was in the opposite direction. Feedback to low-confidence corrects captured more attention than did feedback to high-confidence corrects, $\gamma = .35$, $t(35) = 3.8$, $p < .001$.

Tone Detection and Final Test Accuracy

Two participants corrected all error commissions at retest. For the remaining participants, the mean gamma between tone detection and retest performance was significantly less than zero, $\gamma = -.31$, $t(33) = -3.2$, $p < .005$. This confirmed the

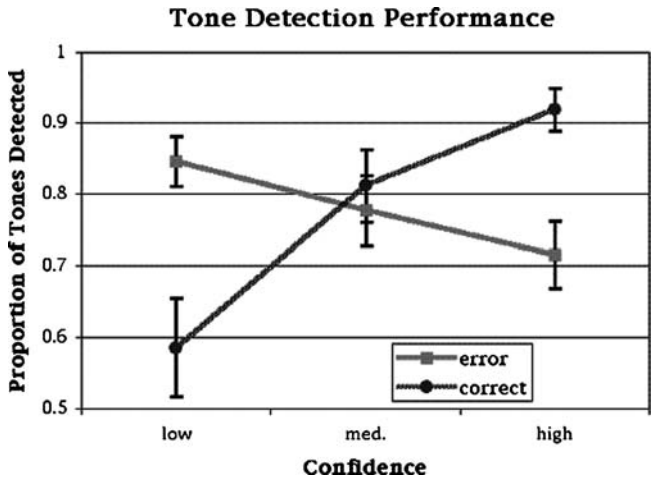

Fig. 1. Experiment 1, proportion of feedback tones detected split by response accuracy and confidence (restricted to trials with a feedback tone presentation), +/− one standard error of the mean. Low confidence is between −50 and −16 (inclusive), medium between −17 and +17 (inclusive), and high is between +17 and +50 (inclusive). Only data from the 25 participants with all 6 types of trials are included

hypothesis that a failure to detect the tone would be associated with better final performance.

Discussion

Tone detection performance was impaired when participants were confronted with corrective feedback to high-confidence errors. Failure to detect a tone during corrective feedback was also associated with improved retest performance. These two results, taken together, suggest that the feedback to high confidence errors captures attention, and that this increased attention improves memory.

One possible problem with Experiment 1 was that the probability of tone presentation varied systematically with trial types of interest. Though one might consider this to have worked against the hypothesis (because participants may have learned to expect tone presentations during feedback to high-confidence errors, and thus detected more of these tones), the possibility exists that participants were using a more complicated strategy. They may have, for example, expected each trial type to have the same probability of tone presentation, and thus under-responded to trial types with higher tone presentation probabilities. To offset this potential confound and to attempt replication of the observed relations between tone detection, metamemory mismatch, and retest accuracy, the probability of tone presentation was equalized across response types in Experiment 2.

Experiment 2

Method

Participants

Forty-five Columbia undergraduates (25 women and 20 men, mean age 23 years) were paid $10 each for participating in the hour-long experiment. The data from six participants were excluded because of aberrant tone detection performance. Four participants failed to do the tone detection task, detecting fewer than 3% of the feedback tones. Two other participants appeared to forget about the tone detection task and then remember it again (and, for one of them, forget about it second time) as the experiment progressed. Their tone detection performance had highly significant autocorrelations of .93 and .79. Thus, whether or not they detected a tone was highly predictive of whether or not they would detect the next tone, which suggests that the variance in their tone detection was largely a function of whether or not they were cognizant of the tone detection task. Participants were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992).

Materials

The same pool of 619 general information questions employed in Experiment 1 was used.

Procedure

The procedure of Experiment 2 was identical to that of Experiment 1 in all respects except the following. The duration of the initial test was increased to 35 minutes, and the duration of the interpolated task was increased to 5 minutes. The experiment was conducted in a room with constant background noise (a fan) to keep participants' hearing from getting more sensitive as the experiment progressed. The tone was presented for a random 75% of the questions and a random 75% of feedback stimuli, regardless of response type.

Results

Each participant's response probability was weighted equally in the calculation of the mean probabilities displayed in Table 3. Though these only include data from participants who had trials in all cells in a given column, the within-participant gamma correlations were computed for all participants with variance in the two measures being correlated.

Basic Data

Participants responded to an average of 29% of the questions correctly at initial test. Participants gave omit responses on an average of 29% of the trials. As in Experiment 1, these trials will not be included in any analyses. Among non-omit responses, participants responded to an average of 42% of questions correctly at first test. Most errors of commission were corrected at retest—average retest accuracy for non-omits was 71%.

Participants' confidence ratings were predictive of initial test accuracy: the mean gamma correlation between confidence and initial test accuracy was .63, $t(38) = 18.5$, $p < .001$.

Tone detection performance was again better when the tone was presented during the question than when it was presented during the feedback—on trials in which participants did not give an omit response, participants detected 80% of the question tones and 72% of the feedback tones, $t(38) = 2.7$, $p < .01$. Any mention of "tones" in the remainder of the results section refers to the tones presented during feedback.

Table 3 Conditional probabilities [with *SEM*] of responses, Experiment 2

P(response confidence)	Given resp. conf.:	P(correct at 1st test)	Given incorrect at 1st test and resp. conf.:	P(correct at retest)
Omit 0.29 [0.03]	*Omit*	- [-]	*Omit*	0.48* [0.03]
Low 0.23 [0.02]	*Low*	0.16 [0.02]	*Low*	0.67* [0.03]
Med. 0.19 [0.02]	*Med.*	0.27 [0.03]	*Med.*	0.71* [0.04]
High 0.29 [0.02]	*High*	0.70 [0.03]	*High*	0.77* [0.03]

*mean from the 38 participants with data in all four cells.

Hyper-correction Effect

There was a significant hypercorrection effect: the mean gamma correlation between confidence in the original error and retest accuracy was significantly positive, $\gamma = .16$, $t(38) = 4.7$, $p < .001$.

Tone Detection and Confidence

As in Experiment 1, increased confidence in an error was associated with impaired tone detection (see Fig. 2). Feedback to a high-confidence error captured more attention than did feedback to a low-confidence error: the mean gamma correlation between error confidence and tone detection for trials on which a tone was presented was $-.21$, $t(38) = -4.2$, $p < .001$.

Feedback to low-confidence corrects also captured more attention than did feedback to high-confidence corrects. One participant's data had to be excluded from the analysis because he or she detected every tone for correct feedback. Data from the remaining participants revealed a significantly positive gamma correlation between confidence in correct responses and tone detection of .38, $t(38) = 5.0$, $p < .001$.

Tone Detection and Final Test Accuracy

As in Experiment 1, the mean gamma between tone detection and retest performance for errors of commission was $-.30$. This correlation was significantly less than zero, $t(38) = -3.6$, $p = .001$. This correlation indicates that people were more likely to be correct on those items on which they missed a tone, than on those on which they had detected a tone. This result provides support for

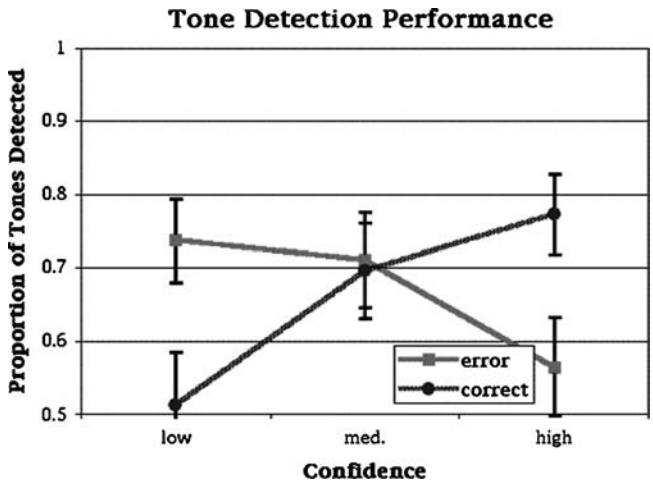

Fig. 2. Experiment 2, proportion of feedback tones detected split by response accuracy and confidence (restricted to trials with a feedback tone presentation), +/− one standard error of the mean. Only data from the 28 participants with all 6 types of trials and variability in tone detection within both incorrect and correct trials are included

the idea that failure to detect the tones was an indication of attentional capture by the feedback.

Discussion

The probability of tone presentation was not based on trial-type in Experiment 2. It replicated the effects found in Experiment 1, namely: (1) commission errors were more likely to be corrected at retest when they were endorsed with higher confidence (the hypercorrection effect), (2) for commission errors, tone detection performance was impaired when people were given corrective feedback to higher confidence errors, relative to lower confidence errors, (3) for correct responses, tone detection performance was impaired when people were given feedback to lower confidence corrects, relative to higher confidence corrects, and (4) for commission errors, tone detection failure was significantly predictive of subsequent memory.

General Discussion

These experiments used a tone detection secondary task to assess the degree to which a participant's attention was captured by feedback to their response. The data from Experiments 1 and 2 demonstrated that corrective feedback on errors on general information questions captures extra attention when those errors were committed with high-confidence. Feedback to low-confidence correct responses also captured disproportionate attention. Both experiments demonstrated that failure to detect the tone was positively related to subsequent retest accuracy on those trials. These two findings provide good support for an attentional factor in the hypercorrection effect.[2]

Conclusion

We often say that we learn from our mistakes. The present studies provide insight into this learning process. Both experiments replicated the finding that errors endorsed with higher confidence are more likely to be corrected at a subsequent retest than are errors endorsed with lower confidence. Two factors appear to contribute to this effect. The first is that questions tapping knowledge for which the subject has more domain familiarity are more likely to elicit high-confidence errors than are questions tapping knowledge in less familiar domains. The feedback to these high domain familiarity questions is more likely to be familiar and, thus, easier to remember. This is evidenced by the correlations of error confidence with question difficulty in Butterfield and Metcalfe (2001b), and the subject familiarity ratings in Butterfield and Mangels (2003).

[2] This memory enhancement for surprising feedback may not override pre-existing memory strength for all populations. The pre-existing memory strength of high-confidence errors may be less mutable for elder subjects, as they are more likely to repeat these errors at retest than are younger subjects. They also evidence a reduced hypercorrection effect relative to younger subjects (Butterfield & Stern, in preparation).

The hypercorrection effect, though, is still robust when familiarity is partialed out–suggesting a second factor. The frontal P3 observed in Butterfield and Mangels' (2003) study has been, in other experiments, associated with novelty or surprise. It was larger for errors that were accorded more confidence and was associated with better memory for the corrected answers at retest, consistent with a postulated connection between the p3 and a feedback loop that bolsters the encoding of the surprising item in memory (as given in Metcalfe, l993b). Both of the present experiments found that when people were considering the feedback to high confidence errors they tended to miss or ignore tones presented in a secondary tone-detection task. These results indicate that their attention had, indeed, been captured by the corrective feedback. Thus, these results support the conclusion that increased attention to high-confidence error feedback, and the resultant enhanced memory encoding, is a critical causal component in the hypercorrection effect.

Acknowledgments This research was supported by National Institute of Mental Health Grant RO1MH60637. We thank Lisa Son, Nate Kornell, Bridgid Finn and the members of Metalab for their help and comments.

References

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review, 80*, 417–438.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*, 97–123.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*, 186–197.

Arnell, K. M. (2001). Cross-modal interactions in dual-task paradigms. In K. Shapiro (Ed.), *The limits of attention: Temporal constraints in human information processing.* (pp. 141–177). London, England: Oxford University Press.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory.* Vol. 2 (pp. 89–195). New York: Academic Press.

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review, 5*, 1–21.

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

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

Butterfield, B., & Metcalfe, J. (2001a). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1491–1494.

Butterfield, B., & Metcalfe, J. (2001b). *Updating the egregious: The relationship between confidence and error correction.* Poster presented at the annual meeting of the American Psychological Society.

Butterfield, B., & Stern, Y. (2006, in preparation). A reduction of the hypercorrection effect in elders.

Courchesne, E., Hillyard, S. A., & Galambos, R. (1975). Stimulus novelty, task relevance, and the visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology, 39*, 131–143.

Crombez, G., Eccleston, C., Baeyens, F., & Eelen, P. (1996). The disruptive nature of pain: An experimental investigation. *Behaviour Research & Therapy, 34*, 911–918.

Dodson, C. S., Koutstaal, W., & Schacter, D. L. (2000). Escape from illusion: Reducing false memories. *Trends in Cognitive Sciences, 4*(10), 391–397.

Gallo, D. A., McDermott, K. B., Percer, J. M., & Roediger, H. L. I. (2001). Modality effects in false

recall and false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 339–353.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.
Glenberg, A. M.,Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*, 597–602.
Goldstein, A., Spencer, K. M., & Donchin, E. (2002). The influence of stimulus deviance and novelty on the P300 and Novelty P3. *Psychophysiology, 39*, 781–790.
Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers, 16*, 96–101.
Hunt, R. R., & Lamb, C. A. (2001). What causes the isolation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1359–1366.
Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General, 188*, 115–125.
Jolicoeur, P. (1999). Restricted attentional capacity between sensory modalities. *Psychonomic Bulletin & Review, 6*, 87–92.
Kelley, C. M., & Lindsay, S. D. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 34*, 1–24.
Knight, R. T. (1984). Decreased response to novel stimuli after prefrontal lesions in man. *Electroencephalography and Clinical Neurophysiology, 59*, 9–20.
Loftus, E. F. (1979). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.
McGeoch, J. A. (1942). *The psychology of human learning.* New York: Longmans.
Madden, D. J. (1986). Adult age differences in the attentional capacity demands of visual search. *Cognitive Development. 1*, 335–363.
Melton, A. W., & Irwin, J. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology, 53*, 173–203.
Metcalfe, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*, 145–160.
Metcalfe, J. (1993a). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3–22.
Metcalfe, J. (l993b). Monitoring and gain control in an episodic memory model: Relation to P300 event-related potentials. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (pp. 327–354). Hillsdale, NJ: Lawrence Erlbaum Associates.
Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627–661.
Müller, G. E., & Schumann, F. (1894). Experimentelle Beiträge zur Untersuchung des Gedächtnisses. *Zertschrift für Psychologie, 6*, 81–190, 257–339.
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 and Verbal Behavior, 19*, 338–368.
Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56*, 132–143.
Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93–134.
Sokolov E. N. (1963). *Perception and the conditioned reflex*. London: Pergamon Press.
Tulving, E. & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review, 2*, 387–390.
Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582–600.
Von Restorff, H. (1933). Uber die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschun*, *18*, 299–342.
Webb, L. W. (1917). Transfer of training and retroaction: A comparative study. *Psychological Monographs, 24*(3), 1–90.
Wright, D. B, & Loftus, E. F. (1998). How misinformation alters memories. *Journal of Experimental Child Psychology, 71*, 55–164.