Journal of Experimental Psychology:
Learning, Memory, and Cognition
2011, Vol. 37, No. 2, 437–448

0278-7393/11/$12.00 DOI: 10.1037/a0021962

People's Hypercorrection of High-Confidence Errors: Did They Know It All Along?

Janet Metcalfe and Bridgid Finn
Columbia University

This study investigated the "knew it all along" explanation of the hypercorrection effect. The hypercorrection effect refers to the finding that when people are given corrective feedback, errors that are committed with high confidence are easier to correct than low-confidence errors. Experiment 1 showed that people were more likely to claim that they knew it all along when they were given the answers to high-confidence errors as compared with low-confidence errors. Experiments 2 and 3 investigated whether people really did know the correct answers before being told or whether the claim in Experiment 1 was mere hindsight bias. Experiment 2 showed that (a) participants were more likely to choose the correct answer in a 2nd guess multiple-choice test when they had expressed an error with high rather than low confidence and (b) that they were more likely to generate the correct answers to high-confidence as compared with low-confidence errors after being told they were wrong and to try again. Experiment 3 showed that (c) people were more likely to produce the correct answer when given a 2-letter cue to high- rather than low-confidence errors and that (d) when feedback was scaffolded by presenting the target letters 1 by 1, people needed fewer such letter prompts to reach the correct answers when they had committed high- rather than low-confidence errors. These results converge on the conclusion that when people said that they knew it all along, they were right. This knowledge, no doubt, contributes to why they are able to correct those high-confidence errors so easily.

Keywords: metacognition, feedback, error correction, memory

It is generally agreed that providing corrective feedback to a person who has made an error is an effective means of rectifying that error (Anderson, Kulhavy, & Andre, 1972; Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008; Kang, McDermott, & Roediger, 2007; Kulhavy, 1977; Lhyle & Kulhavy, 1987; Metcalfe & Kornell, 2007; Metcalfe, Kornell, & Son, 2007; Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005). How beneficial the feedback will be, however, appears to be modulated by people's confidence in their errors. In contrast to what might be expected, errors that are endorsed with higher confidence are more likely to be corrected on a final test than are errors endorsed with lower confidence (Butterfield & Mangels, 2003; Butterfield & Metcalfe, 2001, 2006; Fazio & Marsh, 2009; Kulhavy & Stock, 1989; Kulhavy, Yekovich, & Dyer, 1976). This result is surprising because it indicates that people most easily overwrite the responses that they hold most strongly and correct the erroneous beliefs that are most deeply entrenched, although it seems intuitive that these beliefs and habits should be hardest to change.

In the standard paradigm used to investigate this hypercorrection phenomenon (see, e.g., Butterfield & Metcalfe, 2001), participants were asked to generate the answers to general information questions and to rate their confidence in the correctness of each answer they produced.[1] They were then given the correct answer. At later test, it was found that people were more likely to respond correctly to the questions that had produced high- rather than low-confidence errors. This result occurred despite the fact that most theoretical perspectives on memory and its relation to confidence (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991; Hollingworth, 1913, who discussed Strong's 1912 confidence judgment–memory experiment; Koriat, l997; Koriat, Goldsmith, & Pansky, 2000; Murdock, l974) indicate that responses that are made with high confidence are those in which the person believes most strongly or that are the strongest in memory (e.g., Ebbesen & Rienick, 1998; Tulving & Thomson, 1971). As such, they should be most easily accessible and most resistant to interference. Cer-

Janet Metcalfe and Bridgid Finn, Department of Psychology, Columbia University.

Bridgid Finn is now at Department of Psychology, Washington University in St. Louis.

This research was supported by National Institute of Mental Health Grant RO1-MH60637, Grant R305H060161 from the Institute of Educational Sciences of the Department of Education, and Grant 220020166 from the James S. McDonnell Foundation. The authors are wholly responsible for the content of this article. We thank Barbie Huesler and Frederica Stahl for their help.

Correspondence concerning this article should be addressed to Janet Metcalfe, Department of Psychology, Columbia University, New York, NY 10027. E-mail: jm348@columbia.edu

[1] The hypercorrection effect has even been observed in one experiment in which the first test was multiple choice and the correct alternative was among the alternatives when the wrong alternative was chosen with high confidence. This result is difficult to explain. However, although one experiment showed hypercorrection even under these conditions, a second experiment revealed no advantage to high-confidence errors when the correct alternative was explicitly rejected—in favor of a mistake—on the first test (Butler & Roediger, 2008).

tainly, in all data presented to date on the hypercorrection effect (including in this article), the overall correlation between confidence and correctness is very high. The responses, on average, in which people are highly confident are nearly always correct and are thought to be the strongest, most entrenched responses associated with their respective cues. High-confidence errors should, therefore, be difficult rather than easy to overwrite. Nevertheless, the empirical data indicate that these errors are the easiest, rather than the most difficult, to change.

Two non–mutually exclusive explanations for this phenomenon have been proposed. The first is an attentional explanation. The idea is that when people are wrong with high confidence, they are surprised (and perhaps embarrassed) and they therefore rally their attentional resources to learn the correct item. Several lines of research (Butterfield & Mangels, 2003; Butterfield & Metcalfe, 2001, 2006; Fazio & Marsh, 2009) offer support for this explanation. For example, Butterfield and Metcalfe (2006) showed that people are more likely to miss detecting a soft tone in a concurrent task when it is presented in the interval during which visual feedback is given to a high-confidence error rather than a low-confidence error. Presumably this result obtains because people's attention is captured by the feedback in the high-confidence error condition and they have less in reserve with which to detect the tone. Butterfield and Mangels (2003) investigated the hypercorrection effect by looking for a p300 or *late positivity* event-related potential, a deflection that is thought by most researchers to be an indication of enhanced attention to a novel stimulus. This late positivity is associated with enhanced memory (Paller & Wagner, 2002; Paller, Kutas, & Mayes, l987). Butterfield and Mangels (2003) observed a p300 event-related potential associated with the presentation of corrective feedback to incorrect responses. Of critical importance was the fact that its magnitude was directly related to the person's original confidence in the error. Feedback to high-confidence errors produced a larger p300 than did feedback to low-confidence errors. The authors interpreted this finding as indicating that people were paying more attention to the feedback to high-confidence errors than the feedback to low-confidence errors. Finally, Fazio and Marsh (2009) found that memory for contextual aspects, such as the surface appearance, of the corrections to high-confidence errors was enhanced, a finding that they attributed to increased attention to these corrections. While acknowledging the importance of attentional factors in this phenomenon, we here focus on the second explanation.

The second explanation for which there is some preliminary support is a familiarity account. The general idea is that there may be either systematic differences in the characteristics of general information questions and their answers that are related to high-confidence errors or individually based differences in the participant's own familiarity with the domains of his or her own high- as compared with low-confidence errors. Because of a greater familiarity for high- as compared with low-confidence error domains, the correct answer may already have been partially learned in the more familiar domains of the high-confidence errors and less well-learned or not learned at all in the less familiar domains of the low-confidence errors. Consider a typical high-confidence error such as answering "Toronto" to the question "What is the capital of Canada?" When the person is told that actually the capital is Ottawa, he or she may find this response easy to learn because this person already knows that Ottawa is a city in Canada, and he or she might even have known that it is the capital of Canada, had he or she really thought about it. He or she might have known it all along but made a slip in producing, instead, the more familiar but incorrect response of "Toronto." Now consider a hypothetical low-confidence error, such as saying that Bamako is the capital of Burundi. When the person is told that actually Bujambura is the capital of Burundi, he or she is probably not very familiar with Bujambura (and maybe not be familiar with Burundi, either), so more new learning is needed. He or she did not know it all along.

There is some evidence to support this familiarity-based explanation. Butterfield and Metcalfe (2006) reanalyzed the data from their original 2001 article. The general information questions used in the 2001 article were taken from Nelson and Narens's (1980) article, which had presented the normative values of a correct response for each question in the set. Thus, Butterfield and Metcalfe (2006) were able to assess the normative probability of a correct response for errors committed at various levels of confidence. These were .19 for errors committed with low confidence, .18 for errors committed with medium confidence, and .28 for errors committed with high confidence. This difference in the characteristics of the questions as a function of people's confidence in their errors was significant. They also found that the normative ease of questions answered incorrectly at first test was significantly correlated with later correct recall on the second, postfeedback test. This familiarity or prior learning effect did not account for the whole hypercorrection effect. When they partialed out normative difficulty, there was still a significant residual contributing to the hypercorrection effect. But, even though it was not the whole story, familiarity was implicated.

Butterfield and Mangels (2003) asked their participants for subjective familiarity ratings following the presentation of the correct answer. They found that the correct answers presented following high-confidence errors were retrospectively rated as more familiar than were correct answers following low-confidence errors. They also found, in their event-related potential data, an inferior-temporal negativity occurring 300–600 ms after presentation of the correct answer that was sensitive to subsequent memory performance at both immediate and delayed retests, but only for answers containing familiar semantic information. They suggested that this negativity might reflect processes involved in the formation of an association between the question and preexisting semantic information. These results on the familiarity of the answers to high-confidence errors suggest that people might be more likely to have the answers in their semantic memory.

Here we test the possibility that when people make high-confidence errors they actually know something about the correct answer, and more than they know about the correct answers to low-confidence errors. The question we ask in the first experiment is, Do people exhibiting the hypercorrection effect assert, when the correct answers to high-confidence errors are presented, that they knew it all along?

Experiment 1

In this experiment, college students were queried with general information questions and provided their answers, giving their confidence in each response, until they made 15 errors. After each error and confidence judgment, they were given corrective feedback followed immediately by the question of whether they had known the answer all along. Then a final cued recall test was given. We expected the participants to exhibit the hypercorrection effect, replicating previous research. We also hypothesized that people might, in response to the correct answer, say that they "knew it all along" disproportionately to high-confidence errors rather than to low-confidence errors.

Method

Participants. The participants were 25 undergraduates at Columbia University and Barnard College. They participated for course credit or cash ($10/hr). All participants were treated in accordance with American Psychological Association ethical guidelines in this experiment and the experiments that follow.

Materials. Participants were asked general information questions from a pool of 191 general information questions, which had been taken from the set of Nelson and Narens (1980). A number of questions that were in the original pool were no longer relevant or correct and were eliminated from the pool. Examples of questions used in Experiment 1 were "What is the name of the unit of measure that refers to a six-foot depth of water?" (answer: fathom) or "What is the name of the French author who wrote *The Stranger*?" (answer: Albert Camus).

Procedure. At the beginning of the experiment, participants were instructed that they would be answering general information questions and indicating how sure they were of their answers, that they would then be given the correct answers, and that, following feedback to errors, they would be asked if they had known the answers all along. They were presented with general information questions one at a time and instructed to enter their answers into the blank slot on the computer. They provided their confidence rating concerning the correctness of the answer, using a horizontal slider on the computer that ranged from *very unsure* on the left end to *very sure* on the right end. The slider bar was anchored to the middle of the scale at the onset of each question, and the individual had to move it away from that center rating to have the confidence response register. Confidence ratings were coded along a scale from 0 to 1.00, with 0 indicating a selection of the lowest limit of the slider, at the *very unsure* end, and 1.00 indicating a selection of the highest limit, at the *very sure* end. In the analyses that follow, we bifurcated the rating scale into high confidence and low confidence for above and below .50; we also analyzed high or low confidence on the basis of each participant's median confidence rating, and we used the numerical values of the ratings.

When the participant's answer was correct, a chime sounded and the next general information question was presented. If his or her answer was incorrect, the correct answer was presented on the screen, and the participant was asked to indicate whether he or she knew that answer all along using a second slider anchored, initially, to the center position, which ranged from *That's new to me* on the far left end to *I actually knew it all along* on the far right end. This process continued until participants had answered 15 items incorrectly. These 15 items became the items over which the person's original confidence in his or her errors, as well as the person's "knew it all along" judgments, were computed. Once 15 incorrect items had been accumulated, the program randomized those 15 originally incorrect responses and retested each for a final cued recall test. At the end of the experiment, all participants were thanked and debriefed.

Results

We hand-checked each response of every participant to every question, in each of the three experiments presented here, to be sure that no response was ever counted as incorrect because of a spelling or typing mistake. Any such possibility was eliminated from the data analyzed.

Basic data. On average, participants answered 20.48 (SE = 0.54) questions before they reached the 15 incorrect-answer criterion. Participants' initial confidence in their answers, including both correct and incorrect answers, was .36 (SE = .02). These confidence ratings were predictive of initial test performance. The mean gamma correlation (γ) between initial confidence ratings and initial recall performance, computed on each participant and then averaged over participants, was .83 (SE = .03), which was significantly different from zero, $t(24) = 26.92$, $p < .001$. (In subsequent analyses, we were sometimes unable to report a gamma correlation for some participants because some got everything right or everything wrong or had too many ties and the statistic could not be computed. Thus, degrees of freedom listed for gamma correlations may differ from the total number of participants used in the experiment.) For the items that were answered incorrectly on the initial test, mean prefeedback confidence in the incorrect responses was .22 (SE = .02). Mean postfeedback recall performance on the final test was .76 (SE = .03).

The hypercorrection effect. A hypercorrection effect would be in evidence if high-confidence errors were more likely to be corrected on the final test than errors endorsed with lower confidence. Results showed a significant hypercorrection effect: The mean gamma correlation between confidence in the original error and retest accuracy was.40 (SE = .10), which was significantly greater than zero, $t(22) = 4.04$, $p < .01$.

Knew-it-all-along judgments. Participants' mean knew-it-all-along judgment was .28 (SE = .03). The question of interest, concerning the possibility that high-confidence errors were more likely to be thought to have been known all along, was whether the corrections to the errors that had been endorsed with high confidence were given higher knew-it-all-along judgments than were the corrections to the errors endorsed with low confidence. The correlation between knew-it-all-along judgments and confidence in the original error was $\gamma = .30$, $SE = .05$, $t(24) = 5.84$, $p < .001$, and $\tau_B = .25$, $SE = .04$, when computed with Kendall's τ_B, $t(24) = 5.73$, $p < .001$. A further assessment showed that the mean knew-it-all-along judgment to the corrective feedback was higher for high-confidence errors than for low-confidence errors. When judgments of .50 and greater were classified as high confidence and judgments of .49 or lower were classified as low confidence, the results were significant (for low confidence, $M = .25$, SE = .03; for high confidence, $M = .49$, $SE = .07$), $t(22) = 3.45$, $p <$ .01. Results were also significant when low and high confidence were assessed by using each participant's median overall confidence rating as the split point (for low confidence, $M = .22$, SE = .03; for high confidence, $M = .46$, $SE = .05$), $t(24) = 4.86$, $p <$ .001.

Finally, we computed the correlation between knew it all along judgments and final test performance. The gamma correlation was .50 (SE = .09), which was significantly different from zero, $t(22) = 5.57$, $p < .001$. Thus, when they said they knew it all along, people were more likely to get the answer correct later.

Mediation analyses. To examine the relationships among confidence, knew-it-all-along judgments, and final test performance further, we used a mediational model in which we assessed whether the effect of confidence on final test performance was mediated by the knew-it–all-along judgments. Following the technique recommended by Baron and Kenny (1986), we found evidence of mediation of the impact of confidence on final test performance. There was a significant effect of confidence judgments on final test performance, $\beta = .18$, $t(362) = 3.45$, $p < .01$, and on knew-it-all-along judgments, $\beta = .27$, $t(362) = 5.43$, $p < .001$. There was also a significant direct effect between knew-it-all-along judgments and final test performance, $\beta = .24$, $t(362) = 4.77$, $p < .001$. As illustrated in Figure 1, when both confidence judgments and knew-it-all-along judgments were included as predictors in the regression equation, knew-it-all-along judgments still predicted final test performance, $\beta = .21$, $t(361) = 3.98$, $p < .001$, as did confidence judgments, $\beta = .12$, $t(361) = 2.30$, $p < .05$. The decrease in the direct effect of confidence on final test performance was statistically significant, as measured by a Sobel test, $z = 2.67$, $p < .01$, indicating that the effect of confidence on final test performance was at least partially mediated by knew-it-all-along judgments.

Discussion

The results indicated that, at least some of the time, people believed they knew the correct answers all along. Furthermore, they were more likely to make this claim after receiving feedback to high- as compared with low-confidence errors. If they actually were more likely to know the answers to high- as compared with low-confidence errors, this would provide strong support for the familiarity explanation of the hypercorrection effect. This statement of belief, however, was assessed only after they had seen the correct answers. Whether their claim that they knew the answers all along was a pure hindsight bias or whether it was an indication that they really did know something more than the incorrect answers at the time of making them, before getting corrective feedback, is the issue that is investigated in the second and third experiments.

Experiment 2

Experiment 2 examined the question of whether people showed any hint of knowing the answers to high-confidence errors all along before getting explicit feedback about those correct answers. If participants did have the correct information stored prior to receiving the feedback, then, once alerted to the error, they might have been able to provide the correct answer. The knew-it-all-along claim, though, was selective to high-confidence errors. Accordingly, we would expect that people would be able to produce the answers selectively after a high-confidence error but not after a low-confidence error. The idea that people might be able to produce those answers at least some of the time is convergent on the notion that one reason people make high-confidence errors, despite saying shortly after error commission that they knew it all along, was that the initial incorrect response had been impulsive.

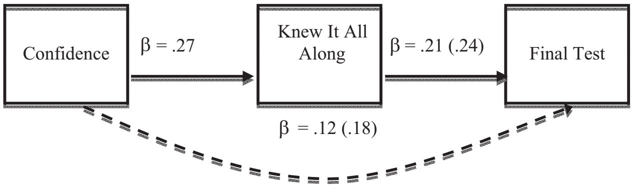

Figure 1. Mediational model of direct and indirect effects of confidence and knew-it-all-along judgments on final test performance, Experiment 1. Values in parentheses indicate the direct effect before the mediator was included in the analysis. All values were significant at $p < .05$.

Kornell and Metcalfe (2006) observed that when people are in tip of the tongue (TOT) states—states in which their metacognitive feelings of potentially knowing the answer are high—they have a good chance of retrieving the target later when given more time to do so (see Schwartz, 2002). Furthermore, people are able to retrieve the correct answer when given more time or an additional retrieval opportunity, even when they were initially in what is called a blocked TOT state in which an incorrect answer has come to mind. The difference between a blocked TOT state and the kind of high-confidence error state observed in the preceding experiment may primarily revolve around whether people know or do not know that the answer that came to mind was incorrect. But it is possible that in the high-confidence error case, as in the blocked TOT case, if people are given more time and put in more effort, the correct answer might be recallable. Furthermore, it has been shown (O'Neill & Douglas, 1996) that boys who are impulsive spend less time trying to recall, and recall less as a result, than do typical nonimpulsive boys. Finally, reminiscence effects (Payne, 1987) suggest that additional retrieval effort may result in recall that is not apparent on the first attempt. These lines of research suggest that, given more time or further effort, high-confidence errors might be correctible (although note that Dijksterhuis, 2007, showed that longer decision time resulted in worse rather than better decisions, contradicting the idea that more time necessarily leads to better results).

No experiments on the hypercorrection effect have included a condition in which no feedback is given and in which people are simply given the final test. The first possibility that we explored was that even without feedback, people might be more likely to correct their high-confidence errors than their low-confidence errors on a subsequent test. Selective self-correction following no corrective feedback would constitute evidence that people might, indeed, have known the answers all along.

If participants did not spontaneously produce the correct answers on the final test, they might have been able to do so—again, taking seriously the idea that they might have known the answers all along at time of the first test—if we told them that they had been incorrect and asked them to try again to generate the correct response before going on. Perhaps, at least some of the time, they could immediately generate the correct answer if we slowed them down and asked them to think again. Because the knew-it-all-along responses were selective to high-confidence errors, we expected that people would be more likely to generate the correct answer to questions that had evoked high- rather than low-confidence errors.

If participants were unable to generate the correct response themselves, they might still have been able to correctly recognize the answer from a list of alternatives, especially when they had made a high- rather than low-confidence error. This too would suggest that their knew-it-all-along ratings had had some basis in fact and were not purely a hindsight bias. These three conditions—

the *no-feedback* condition, the *generate* condition, and the *multiple-choice* condition—were all contrasted with the *standard feedback* condition in Experiment 2.

The fact that we asked people to make a second guess in our generate and multiple-choice conditions raises the possibility that these conditions might relate to the interesting work of Mozer, Pashler, and Homaei (2008) and Vul and Pashler (2008) in which they showed that having people make two guesses in a magnitude estimation task resulted in a better estimate of the true magnitude than did having them make only a single guess. This improvement in estimation with multiple sampling is a well-known finding, when the sampling is conducted by having different people make independent guesses about things like the weight of an ox. This finding is often dubbed the "wisdom of crowds." It also occurs, though, with estimates made by a single person who, at different times, takes different samples from memory, suggesting that there may be what the authors called "a crowd within." Although, in our paradigm, we were not asking people to estimate a quantity, such as the weight of an ox; we were asking them to sample their memory more than once. As such, if multiple answers were available, participants might also show the beneficial effects of "a crowd within" and performance might improve, as Mozer et al. and Vul and Pashler found. Our question, however, was whether the space sampled was better with respect to the true answer when participants had made high- rather than low-confidence errors on the first sampling.

It was not necessarily the case that people would be able to show, in any of these three conditions, that they knew the answers all along preferentially for high-confidence errors, as they had so claimed, prior to feedback. Much research has shown that after people learn the outcome to a situation or the answer to a question, they tend to exaggerate their ability to have seen it coming, claiming to have known the answer all along (see Hawkins & Hastie, 1990; Hoffrage & Pohl, 2003; Sanna & Schwartz, 2006). It was possible that the postfeedback knew-it-all-along judgments we obtained in Experiment 1 were simply a demonstration of a classic hindsight bias (Fischhoff, 1975; Wood, 1978) and when a domain was familiar, the feedback was judged to be a part of the person's knowledge set even if it never could have been retrieved. Indeed, Werth and Strack (2003) showed that people were more likely to show a classic hindsight bias when an answer was judged to be highly familiar.

Experiment 2 could potentially reveal either that people had some knowledge of the correct answers all along when they said they did or that the claim that they had made in the first experiment was purely a hindsight bias. If it was a pure bias, then recalling the correct answer might not have been possible even with a second chance on a retest for which no corrective feedback was supplied. Asking the participants to make a second attempt to generate the answer or giving them a multiple-choice test that included the correct answer but did not include their original response could fail to reveal any bias toward the high-confidence errors in the correctness of the second-chance responses. If none of these three methods revealed evidence that people could demonstrate that they had greater knowledge for the correct responses to high-confidence error questions than to low-confidence error questions, then the results would provide support for the idea that the knew-it-all-along judgments that they had made on the previous experiment were the result of pure hindsight bias rather than actual prior knowledge. Such results would also suggest that differential attention to feedback to high- and low-confidence errors was entirely responsible for the hypercorrection effect.

Method

Participants. The participants were 45 undergraduates at Columbia University or Washington University in St. Louis. They received course credit or cash ($10/hr) for participation.

Procedure. Participants answered general information questions until they had reached 36 errors. The questions were randomly drawn from a larger pool of 493 questions taken from Nelson and Narens (1980) as well as a variety of difficult trivia questions that elicit high-confidence errors added since the earlier pool was constructed but that followed the form of Nelson and Narens's general information questions.

The questions were presented one at a time, in a random order. After each response, which, in this experiment, unlike Experiment 1, was forced (i.e., the participant had to provide a response to get the program to continue), the participants made a confidence judgment about the correctness of their response. The answers were computer scored online by an algorithm, developed by Brady Butterfield, that computed proportion of letter overlap and order, assigning each item a value between 0 and 1, where a score of .75 corresponds fairly well to what human scorers would call a minor spelling mistake. The item was counted as correct if the score was .75 or higher; otherwise, it was treated as an error. When the response was incorrect, a low-pitched honk sounded, and one of four within-participants feedback conditions, randomly determined, occurred: (a) standard feedback, (b) no feedback, (c) generation, or (d) multiple choice. Immediately following each feedback treatment, participants were asked "Did you know that all along?" and had to click either a *yes* or *no* button. There were nine replications in each of the four treatment conditions, for 36 errors in all. After the 36th error, participants were given a cued recall test, in which the 36 items that had been answered incorrectly were randomized and each question was presented for the participant to type the correct answer into the computer. Participants had no restrictions on the amount of time they could take to answer each question.

In the standard feedback condition, after indicating their confidence in their incorrect answer, participants were simply told, "Actually, the correct answer is *x*," and the correct answer was presented in the response window on the computer screen. This condition allowed us to ascertain that the basic paradigm from Experiment 1 replicated. In the no-feedback condition, aside from the fact that participants had heard the chime when they were correct and the low sound when incorrect and so would have known that they had been incorrect, no corrective feedback was given, nor was there a chance to come up with a second guess. This condition allowed us to investigate whether, without any corrective feedback at all, people might, on the final test, change their answers to be correct. In the generation condition, the computer told the participants "Please choose another answer. If you do not know the answer, please guess." After they typed in a new response, they made a knew-it-all-along judgment and were not told whether their response was correct. This condition was directed at the possibility that high-confidence errors were the result of impulsiveness and that if people were told that they were incorrect and asked to give another response, they might be able to produce the correct answers. Finally, in the multiple-choice condition, a message stated, “Actually, the answer is one of these 6 options. Please choose one.” A randomized array of six options, including the correct answer, was presented, and participants could choose a new response. The program randomly selected the six options from a set of nine potential options. If the participant’s original error was included in the list of six options first selected by the computer, that option was replaced, randomly, with one of the remaining three options. After selecting their forced-choice response, participants made a knew-it-all-along judgment and then moved on to the next question.

Results

Basic data. On average, participants answered 51.82 (*SE* = 1.44) questions before they reached the 36 incorrect-answer criterion. Participants’ initial confidence in their answers was .38 (*SE* = .02). Their confidence ratings were predictive of their initial test performance. The mean of gamma correlations taken over participants, between initial confidence ratings and initial recall performance, was .81 (*SE* = .02) and was significantly different from zero, $t(44) = 45.59$, $p < .001$. These basic analyses showed no effect of feedback condition with any measure. This was expected because the feedback manipulation followed the initial test and confidence ratings.

Using only the items that were answered incorrectly on the initial test, of course, there was a significant effect of feedback condition on final test performance, $F(3, 132) = 226.70$, mean square error (*MSE*) = .02, $p < .001$, $\eta_p^2 = .84$ (effect size was computed using partial eta squared, here and throughout). The standard feedback condition showed the highest final test performance ($M = .70$, $SE = .03$), followed by the multiple-choice condition ($M = .30$, $SE = .03$), the generate condition ($M = .07$, $SE = .01$), and the no-feedback condition ($M = .04$, $SE = .01$). All feedback conditions were significantly different from zero (all *t*s > 1, all *p*s < .05). All comparisons between the feedback conditions showed significant differences, all *t*s > 1, all *p*s < .05, except for the comparison between the no feedback and generate conditions, which showed a difference that approached significance, $t(44) = 2.48$, $p = .10$ (all pairwise comparisons in this analysis and in the analyses that follow were Bonferroni corrected to the .05 level).

Because only the standard feedback condition appropriately measured the hypercorrection effect, insofar as none of the other conditions provided corrective feedback, the question of whether there was a hypercorrection effect is addressed below, with regard to that condition alone.

Standard feedback condition. All of the basic effects that were seen in Experiment 1 replicated in this experiment. First, there was a hypercorrection effect. Errors that were committed with higher confidence were more likely than errors committed with lower confidence to be corrected, $\gamma = .28$, $SE = .08$, $t(40) = 3.57$, $p < .01$. There was a significant knew-it-all-along effect, such that participants were more likely to say they knew it all along when given the feedback to high-confidence errors than to low-confidence errors, when confidence was split on the basis of the center of the scale (probability of saying they knew it all along|low-confidence error: $M = .10$, $SE = .02$; probability of saying they knew it all along|high-confidence error: $M = .21$, $SE = .05$), $t(34) = 2.11$, $p < .05$. This effect only approached significance when confidence was split on the basis of each participant’s median confidence rating over all responses, however (probability of saying they knew it all along|low-confidence error: $M = .09$, $SE = .02$; probability of saying they knew it all along|high-confidence error: $M = .17$, $SE = .04$), $t(40) = 1.49$, $p = .07$, one-tailed. When assessed with gamma correlations computed between confidence and whether the person said that they knew it all along, the effect was significant, $\gamma = .31$, $SE = .11$, $t(25) = 2.84$, $p < .01$. Finally, participants were more likely to be correct on the final test if they had claimed that they knew the answer all along, as measured by a phi correlation, $r_\varphi = .30$, $t(23) = 9.48$, $p < .001$.

No-feedback condition. There were very few correct answers in the no-feedback condition ($M = .04$, $SE = .01$). For the participants for which it could be computed—but there were few of them, so this null result should be viewed with caution—we found that the correlation between initial confidence and final accuracy was not significantly different from zero, $t(12) = 1.16$, $p > .05$.

Generation condition. First, and germane to the question of whether there was any evidence that people knew the answers all along, there was a significantly greater probability of generating a correct second-guess response after a high-confidence error than when an error had been produced with low confidence. This result was found whether we divided confidence at the center of the scale (probability of generating a correct second guess|low-confidence error: $M = .04$, $SE = .01$; probability of generating a correct second guess|high-confidence error: $M = .18$, $SE = .05$), $t(37) = 2.80$, $p < .01$, by using a median split (probability of generating a correct second guess|low-confidence error: $M = .04$, $SE = .01$; probability of generating a correct second guess|high-confidence error: $M = .15$, $SE = .04$), $t(41) = 2.94$, $p < .01$, or we computed gamma correlations between original confidence and whether a correct second guess was generated, $\gamma = .48$, $SE = .14$, $t(23) = 3.49$, $p < .01$. This effect is shown in Figure 2. Furthermore, items that produced errors, in this condition, that were committed with higher confidence were more likely to produce correct answers on the final test than were items that produced errors committed with lower confidence, $\gamma = .53$, $SE = .12$, $t(22) = 4.43$, $p < .001$.

The probability of producing the correct answer on the final test was higher when the answer had been produced on the second-guess test (final recall|incorrect second guess: $M = .00$, $SE = .00$; final recall|correct second guess: $M = .81$, $SE = .08$), $t(23) = 10.34$, $p < .001$. Final test performance was higher when the original error had been made with high confidence when the split was based on the center of the scale (final recall|low-confidence error: $M = .04$, $SE = .01$; final recall|high-confidence error: $M = .17$, $SE = .05$), $t(30) = 2.39$, $p < .05$, and when the split was based on participants’ median confidence ratings (final recall|low-confidence error: $M = .04$, $SE = .01$; final recall|high-confidence error: $M = .13$, $SE = .03$), $t(41) = 2.56$, $p = .01$. There was no difference between the second-guess performance and the final recall, with the former being .08, $SE = .01$, and the latter being .07, $SE = .01$, $t(44) < 1$, $p > .05$. Note that this analysis used data from all participants, because each had a score, whereas in the conditionalized analysis presented just prior to this result, a number of participants were excluded from the paired comparison because they did not have any high-confidence observations (which is why

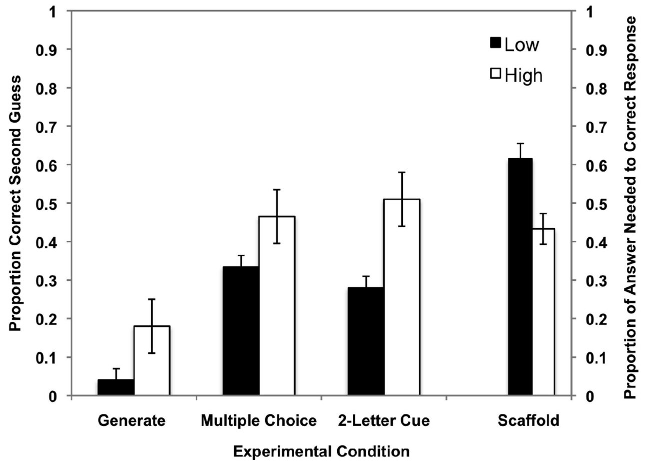

Figure 2. Probability of a correct second guess for low- and high-confidence errors when participants were asked to generate a second response (Experiment 2, far left), when they were asked to choose the correct response in a six-alternative multiple-choice test (Experiment 2, center left), when they were given a two-letter cue (Experiment 3, center right), and the proportion of the word that needed to be revealed to allow participants to produce the correct answer for low- and high-confidence errors (Experiment 3, far right).

there were only 30 degrees of freedom in that *t* test). These results indicate that adults did, at least to some extent, know the answer all along when they said they did—namely, for the high-confidence errors.

Multiple-choice condition. When people had committed a high-confidence error, there was a greater probability of their choosing the correct response on the multiple-choice test than when an error had been produced with low confidence (probability of correct multiple-choice response|low-confidence error: $M = .32$, $SE = .03$; probability of correct multiple-choice response|high-confidence error: $M = .46$, $SE = .08$), $t(31) = 1.91$, $p = .03$, one-tailed, when the split was based on the middle of the scale. This difference did not reach significance, however, when the split was based on participants' median response confidence (probability of correct multiple-choice response|low-confidence error: $M = .34$, $SE = .04$; probability of correct multiple-choice response|high-confidence error: $M = .41$, $SE = .06$), $t(35) = 1.00$, $p > .05$, and was not significant when gamma correlations between original confidence and whether the multiple choice was correct or incorrect were computed, $\gamma = .04$, $SE = .16$, $t < 1$, $p > .05$. Twelve participants (seven in the median confidence split) who had no high-confidence responses at all were included in the gamma analyses, and their inclusion, as well as the increased weighting of the many low-confidence responses in the gamma analysis, may have diluted the effect.

Errors that were committed with higher confidence were more likely to be correctly recalled on the final recall test than were errors committed with lower confidence, $\gamma = .21$, $SE = .09$, $t(38) = 2.40$, $p < .05$. The probability of producing the correct final answer was higher when the correct choice was made on the multiple-choice test than when it had not (probability of correct final recall|incorrect multiple-choice response: $M = .07$, $SE = .02$; probability of correct final recall|correct multiple-choice response: $M = .69$, $SE = .05$), $t(42) = 11.13$, $p < .001$. If the person not only picked the correct answer on the multiple-choice question but also said, after making that choice, that he or she knew it all along, their final test performance was .95, $SE = .04$, as compared with when a person picked the correct answer but said that he or she did not know it all along ($M = .58$, $SE = .08$), $t(25) = 4.51$, $p < .01$. Final recall test performance was higher when the original error had been made with high confidence (final recall|low-confidence error: $M = .27$, $SE = .03$; final recall|high-confidence error: $M = .52$, $SE = .08$), $t(33) = 3.05$, $p < .01$. There was a significant difference between multiple-choice test performance and final recall (for multiple-choice test performance, $M = .35$, $SE = .03$; for final recall, $M = .30$, $SE = .03$), $t(44) = 2.60$, $p < .05$, but this difference is difficult to interpret because there was a .17 guessing probability in the multiple-choice test. When the person did choose the correct answer on a multiple-choice question, the probability of a correct final answer was higher than their final test mean, $t(42) = 9.24$, $p < .01$. These multiple-choice data provide some support (although, perhaps, more equivocal than those provided in the generate condition) for the idea that the participants did know all along at least some of the correct answers to their high-confidence errors.

Discussion

In Experiment 2, once again, the hypercorrection effect result itself was replicated. People also claimed that they knew it all along more frequently to high- than to low-confidence errors, thus confirming the basic findings of Experiment 1. Simply retesting people on the questions to which they had given the wrong answers or even asking participants to generate a second response produced very small (but not zero) final recall benefits. Final proportion recalled when the participant had made an error and received no feedback except yes–no feedback was given was extremely low: .04. When people were explicitly asked to immediately try to generate the correct answer to their errors—knowing only that they were errors—their final recall performance was also very low, although numerically slightly higher: .07. Clearly, giving more complete feedback has a much greater beneficial effect than giving minimal feedback, consistent with the review of the feedback literature on this issue by Pashler et al. (2005).

The question of primary interest in Experiment 2, though, concerned whether people actually knew the answers to high-confidence errors all along. Two lines of evidence from this experiment suggest that they did have some knowledge of the correct answers to high-confidence errors, that is, that their claim that they knew it all along—that was selective to these kinds of errors—had some basis in fact. First, participants were able to generate the answers to errors they had made with high confidence to a greater extent than to errors they had made with low confidence. Second, they were able to pick the correct answers on a second-guess multiple-choice test more frequently to their high-confidence errors than to their low-confidence errors. When they did pick the correct answer and said that they knew it all along, they almost invariably got the answer right on the final test, even without having been given feedback about whether they had been correct on the multiple-choice test.

Thus, the knew-it-all-along claim appeared to be based, partially at least, on the fact that they did know the correct answers preferentially when they had made high- rather than low-confidence errors. The claim did not appear to be a pure hindsight bias. In Experiment 3, we sought additional evidence bearing on this possibility.

Experiment 3

In the third experiment, we sought to test the knew it all along hypothesis using a different method of evoking a response without necessarily providing full corrective feedback. The overarching rationale for Experiment 3 was the same as for Experiment 2: If people did actually know the answers all along more often when they made high-confidence versus low-confidence errors, then they should be more likely to produce the correct answer to those high-confidence questions more easily. In the third experiment, we gave them clues to help them.

There were two conditions in this experiment: the two-letter cue condition and the scaffold feedback condition. In the former condition, after committing an error, people were given the first two letters of the target word and asked to generate the correct answer. In the second condition, participants were given scaffolded feedback (Carpenter & DeLosh, 2006; Finn & Metcalfe, 2010) after having committed an error. First they were asked to make a second guess. If that was not successful, they were given the first letter and again asked to make a guess. Then they were given the second letter, third letter, and so on, with the opportunity to guess the target after each successive letter, until they had correctly guessed the target. After receiving feedback about the target in this way, they were asked whether they knew it all along. The hypothesis with respect to this second condition was that people would need fewer letters for the high-confidence errors than for the low-confidence errors and, once they had seen the entire word, would be more likely to affirm that they knew it all along.

Method

The general method was the same as in Experiment 2, with the changes noted below.

Participants. The participants in this experiment were 24 students at Columbia University or Barnard College, who received course credit for participating.

Procedure. The procedure was the same as that of Experiment 2 except that participants were told that if they were incorrect, half of the time the computer would give them the first two letters of the correct answer and they should try to type in the correct answers from this clue. After being given the first two letters, if they typed in a word that was the correct answer, a ding would sound. Then they would then be asked whether they had known the answer all along. The other half of the time, when they made an incorrect response, they were told that the computer would start by giving them one letter, and they should guess if they could. The computer would continue to give them one letter at a time until they had either produced the correct answer or the entire answer had been revealed. Then they would be asked whether they had known the answer all along.

A total of 72 errors were accumulated, 36 of which were randomly assigned to the two-letter clue condition and 36 of which were assigned to the scaffold feedback condition. Once all 72 errors had received this feedback, a final test was administered in which participants were asked to provide the correct answer to all 72 questions that had just been answered erroneously.

Results

Basic data. On average, participants answered 99.42 (SE = 2.49) questions before they reached the 72 incorrect-answer criterion. Participants' initial confidence in their answers was .34 (SE = .02). Their confidence ratings about their initial answers were predictive of the accuracy of their initial test performance, $\gamma = .83$, $t(23) = 61.56$, $p < .001$.

When we using only the items that were answered incorrectly on the initial test, we found an effect of feedback condition on final test performance, $t(23) = 11.24$, $p < .001$.The scaffold condition produced better results than did the two-letter-cue condition (with proportion recall of .61, $SE = .03$, and .31, $SE = .02$, respectively).

Two-letter cue condition. People were more likely to generate a second guess correctly when given two letters when they had given their erroneous answer high confidence than when they had ascribed low confidence to their answer, assessed by dividing confidence at the center of the scale (probability of correct response to two-letter cue|low-confidence error: $M = .29$, $SE = .02$; probability of correct response to two-letter cue|high-confidence error: $M = .48$, $SE = .06$), $t(23) = 3.32$, $p < .01$, and assessed by splitting the data at participants' median confidence (probability of correct response to two-letter cue|low-confidence error: $M = .27$, $SE = .02$; probability of correct response to two-letter cue|high-confidence error: $M = .42$, $SE = .05$), $t(22) = 3.11$, $p < .01$. The gamma correlation between original confidence and whether a correct second guess was generated was significant, $\gamma = .19$, $SE = .08$, $t(23) = 2.30$, $p < .05$. When participants produced a correct second guess to the two-letter cue, the probability of saying that they knew it all along was .40, $SE = .05$, which was significantly greater than zero, $t(23) = 1.90$, $p < .05$.

On the final test, participants were more likely to be correct on those questions to which they had originally ascribed high rather than low confidence. The proportions correct on the final test were almost identical to those observed during the process of attempting to generate the correct answer from two-letter cues. With a 50–50 split, probability of correct final recall|low-confidence error: $M = .28$, $SE = .02$; probability of correct final recall|high-confidence error: $M = .51$, $SE = .07$), $t(23) = 3.58$, $p > .01$; with a median split, probability of correct final recall|low-confidence error: $M = .26$, $SE = .02$; probability of correct final recall|high-confidence error: $M = .43$, $SE = .05$, $t(22) = 3.55$, $p > .01$ (see Figure 2). The mean gamma correlation between original confidence in the errors and final performance was .22 ($SE = .09$), $t(23) = 2.58$, $p < .05$. These results offer support for the idea that people did, at least to some extent, know the answers all along when they made high-confidence errors.

Scaffold feedback condition. When people had committed a high-confidence error, the number of letters that they needed to produce the correct answer was significantly fewer than the number of letters needed when they had produced an error with low confidence. With a 50–50 split, number of letters required to correct answer|high-confidence error: $M = 2.65$, $SE = 0.27$; number of letters required to correct answer|low-confidence error: $M = 4.15$, $SE = 0.13$, $t(23) = 5.00$, $p < .001$; with a median split, number of letters required to correct answer|high-confidence error:

$M = 3.27$, $SE = 0.21$; number of letters required to correct answer|low-confidence error: $M = 4.31$, $SE = 0.15$, $t(22) = 4.32$, $p < .001$. The average length of the correct answer was longer for low-confidence errors ($M = 6.59$ letters, $SE = 0.06$) than for high-confidence errors ($M = 6.13$ letters, $SE = 0.16$), $t(23) = 2.51$, with a 50–50 split. This difference (for low confidence, $M = 6.56$, $SE = 0.16$; for high confidence, $M = 6.46$, $SE = .10$) was not, however, significant with a median split, $t < 1$, $p > .05$. Because the direction of the difference in the word length between the high- and low-confidence errors favored the relation of shorter, higher frequency words with high-confidence rather than low-confidence errors–as might be expected by the familiarity view of the hypercorrection effect–we also analyzed for the reduced cuing effect by computing, for each response, the proportion of the target word that was necessary to generate a correct response. This analysis was directed at the possibility that the cuing effects might be due to the shorter nature of the words rather than to people knowing the words in the high-confidence condition. Again, though, even when we controlled for the number of letters, the results favored the high-confidence error condition. The analysis showed that a smaller proportion of the word needed to be revealed to allow correct target completion, both when we examined the data with a 50–50 split on confidence (proportion of the word that needed to be revealed for correct response|high-confidence error: $M = .43$, $SE = .04$; proportion of the word that needed to be revealed for correct response|low-confidence error: $M = .61$, $SE = .02$), $t(23) = 3.93$, $p < .001$. The same pattern emerged when we examined the data using a median split (proportion of the word that needed to be revealed for correct response|high-confidence error: $M = .50$, $SE = .03$; proportion of the word that need to be revealed for correct response|low-confidence error: $M = .64$, $SE = .02$), $t(22) = 4.70$, $p < .001$. These results favor the idea that people did know the answers all along preferentially to the high-confidence errors.

The probability of correct responding on the final test was plotted as a function of the proportion of letters that had to be revealed, as is shown in Figure 3. Proportions were grouped into bins of 25%. Correct final performance was greater for answers that required fewer rather than more letters to be revealed, $F(3, 69) = 33.55$, $MSE = .03$, $p < .001$, $\eta_p^2 = .59$, suggesting that the knew-it-all-along factor was important for later correct responding.

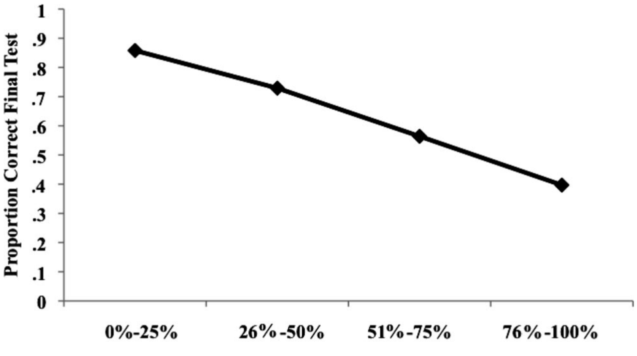

Figure 3. Probability of correct final recall as a function of the proportion of letters revealed in the scaffold condition, Experiment 3.

Because complete feedback was given in this condition—although in a piecemeal way—we were able to look at whether participants claimed that they knew the answers all along for the high- as compared with the low-confidence errors. They did: The proportion of knew-it-all-along responses, using a 50–50 split, was .45 ($SE = .06$) for the high-confidence errors, as compared with .19 ($SE = .02$) for the low-confidence errors, $t(23) = 4.14$, $p < .001$. The proportions were .36 ($SE = .04$) and .17 ($SE = .03$), respectively, $t(22) = 5.60$, $p < .001$, when we used a median split. Finally, there was a hypercorrection effect, such that performance on the final test was better for high- than low-confidence errors. The mean gamma correlation between original confidence and final performance was .17 ($SE = .07$), $t(23) = 2.50$, $p < .05$. Additionally, in this condition, we had sufficient responses per participant to allow us to plot the relation of original confidence in the error and final performance, dividing the data into confidence quartiles, as is shown in Figure 4.

Additional analysis of item characteristics of high- and low-confidence errors. Finally, to examine the possible effects of the characteristics of the items themselves and whether there were item differences that distinguished high- from low-confidence errors, we conducted a latent semantic analysis (LSA; Landauer & Dumais, 1997, and see http://cwl-projects.cogsci.rpi.edu/msr/) to examine the association strength between the error that was generated and the target item when it was a high-confidence error and when it was a low-confidence error. To do so, the LSA value of every error–target pair was determined, and then these pairs were grouped into low-confidence pairs and high-confidence pairs, on a participant-by-participant basis, for analysis. If an error was not a word or could not be found in the database, it was eliminated from this analysis. Although we report each contrast separately, in every case, the associative relation of the low-confidence error to the target was lower than was the associative relation of the high-confidence errors to the target. This was found when the confidence split was based on the midpoint of the scale and collapsed over all three experiments (for low-confidence errors, $M_{\text{LSA error–target}} = .20$, $SE = .01$; for high-confidence errors, $M_{\text{LSA error–target}} = .31$, $SE = .01$), $t(89) = 7.35$, $p < .001$. The same pattern resulted when a participant-based median split was used (for low-confidence errors, $M_{\text{LSA error–target}} = .19$, $SE = .01$; for high-confidence errors, $M_{\text{LSA error–target}} = .30$, $SE = .01$), $t(90) = 8.98$, $p < .001$. The Experiment 1 means were, for midsplit low–confidence errors, $M_{\text{LSA error–target}} = .25$, $SE = .02$, and for midsplit high-confidence errors, $M_{\text{LSA error–target}} = .37$, $SE = .04$, $t(21) = 3.37$, $p < .01$; for median split low-confidence errors, $M_{\text{LSA error–target}} = .24$, $SE = .03$, and for median split high-confidence errors, $M_{\text{LSA error–target}} = .36$, $SE = .03$, $t(23) = 3.14$, $p < .01$. The Experiment 2 means were, for midsplit low-confidence errors, $M_{\text{LSA error–target}} = .18$, $SE = .01$, and for midsplit high-confidence errors, $M_{\text{LSA error–target}} = .29$, $SE = .02$, $t(43) = 5.77$, $p < .001$; for median split low-confidence errors, $M_{\text{LSA error–target}} = .16$, $SE = .01$, and for median split high-confidence errors, $M_{\text{LSA error–target}} = .28$, $SE = .01$, $t(43) = 8.18$, $p < .001$. The Experiment 3 means were, for midsplit low-confidence errors, $M_{\text{LSA error–target}} = .21$, $SE = .01$, and for midsplit high-confidence errors, $M_{\text{LSA error–target}} = .29$, $SE = .02$, $t(23) = 3.42$, $p < .01$; and for median split low-confidence errors, $M_{\text{LSA error–target}} = .18$, $SE = .01$, and for median split high-confidence errors, $M_{\text{LSA error–target}} = .28$,

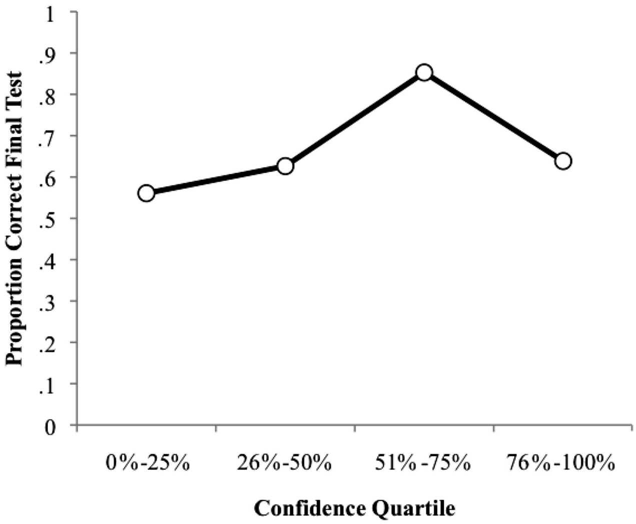

Figure 4. Probability of correct final recall in the scaffold condition as a function of confidence in original error, Experiment 3.

SE = .01, *t*(22) = 7.17, *p* < .001. These differences, like the differences in the normative probability correct for high- versus low-confidence errors reported by Butterfield and Metcalfe (2006), indicate that there were important differences in the characteristics of the items that evoke high- versus low-confidence errors and provide additional support for the familiarity explanation of the hypercorrection effect.

General Discussion

These experiments replicated the finding (Butterfield & Metcalfe, 2001) that subjective confidence in one's errors plays a role in which errors are most likely to be amended: Errors that are endorsed with high confidence are hypercorrected after corrective feedback. We do not dispute our own finding and those of others (Butterfield & Mangels, 2003; Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009) that when a person makes a high-confidence error, he or she gives the corrective feedback to that error extra attention. This attentional boost undoubtedly contributes to the hypercorrection of those errors.

However, as the present results demonstrate, increased attention to feedback after a high-confidence error does not appear to be the whole story behind the hypercorrection effect. People claimed, after receiving the corrective feedback following high-confidence errors, that they knew it all along. When we investigated whether the knew-it-all-along claim had any basis in fact, we found that it did. People were able to selectively generate the correct answers to high-confidence errors, as compared with low-confidence errors, when they were asked to try again. They did this rarely, but they did it more frequently for high-confidence errors than for low-confidence errors (to which they almost never generated a correct second response). They were also sometimes able to select the correct answer in a multiple-choice test, following high-confidence erroneous responses, without first having had corrective feedback. They were more likely to produce the correct answer after seeing the minimal cue of the first two letters of the correct answer, if the error had been made with high rather than low confidence. And, finally, when we gave participants scaffolded cues, such that they got one letter at a time of the answer until they were able to produce the answer, they needed fewer such cues to produce the correct responses to high-confidence errors than to low-confidence errors. Thus, the selective claim that people had made after having been given the correct responses to high-confidence errors—that they knew the answers all along—appears not to be merely a hindsight bias. It appears to have some basis in fact.

The finding that adults did sometimes have knowledge of the correct answer even when they committed a high-confidence error and that they are easily able to correct high-confidence errors given standard corrective feedback (although they almost never corrected them given no feedback) suggests that those items that evoke high-confidence errors may be inside what we have previously called the person's region of proximal learning (RPL; Kornell & Metcalfe, 2006; Metcalfe, 2002, 2009; Metcalfe & Kornell, 2003, 2005). The individual's RPL is thought to comprise materials that are almost but not quite learned and that will benefit the most from additional study opportunities. The materials in the person's RPL are thought to not be difficult to learn, but without further study (or, in this case, feedback, which provides an excellent study opportunity), the individual may fail to learn these items. A small investment of effort has large payoffs. Our no-feedback condition attests to the fact that without such an additional study opportunity, final performance on errors remained at roughly a 4% level of performance. Errors are not spontaneously self-correcting. Given a moment of standard corrective feedback, memory performance climbed to 70% or more. High-confidence errors given corrective feedback reached 82%. Thus, all items benefited from feedback. But high-confidence errors, those items that a priori might have been thought to be impervious to correction, were the easiest to correct both because of the extra attention allocated to the corrective feedback and because people appear—according to the present results—to know quite a bit about those answers all along.

In practical terms, it would appear that encouraging students to generate their own responses, thereby reaping the benefit of the generation or testing effect (Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel & Fisher, 1991; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006a, 2006b; Slamecka & Graf, l978) is likely to be highly beneficial to their learning. The only potential problem with doing so had been the possibility that the errors that people make, in such a generation or test procedure, might prove to be problematic. If making an error entrenched that error in memory and made it harder to learn the correct answer, one might rationally opt not to encourage generation and testing because of the inevitable errors that it entails. This logic would seem to apply most pointedly to the errors that the person believed in most strongly, namely, their high-confidence errors. The present results argue strongly against this rationale for avoiding having students generate the answers. The very high levels of recall on what had earlier been errors, once those errors have been given corrective feedback, indicates that the commission of errors does not harm eventual memory performance, at least in normal college students. Some caution should be exercised in this claim that errors may have few long-term detrimental effects because our participants were typical college students. The commission of errors may have a more detrimental effect in people with memory disorders (Baddeley & Wilson, 1994; Glisky, Schacter, & Tulving, 1986) and perhaps in people with learning disorders. More research on these special populations is needed to determine the limitations and boundary conditions of the conclusions we reach here concerning the effects of errors.

However, with normal college students, the avoidance of generation or testing procedures, which are otherwise beneficial for memory, would not seem to be justified on the grounds that such procedures inevitably result in errors and the commission of errors poses problems for later memory. Indeed, with just a few moments of corrective feedback, the errors themselves are not recommitted but rather are corrected at a very high rate. High-confidence errors are particularly easy, not difficult, to correct. Furthermore, as Butterfield and Metcalfe (2006) showed, the correct answers to high-confidence errors are maintained over time. The one obvious caveat to the idea that committing errors does not harm future memory with typical college students is that those errors do need to be corrected. Feedback is essential. Metcalfe et al. (2009) have shown that the errors do not necessarily have to be corrected immediately. However, as we have shown here, without correction, little or no remediation can be expected.

References

Anderson, R. C., Kulhavy, R. M., & Andre, T. (1972). Conditions under which feedback facilitates learning from programmed lessons. *Journal of Educational Psychology, 63,* 186–188. doi:10.1037/h0032653

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia, 32,* 53–68. doi:10.1016/0028-3932(94)90068-X

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182. doi:10.1037/0022-3514.51.6.1173

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918–928. doi:10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527. doi:10.1080/09541440701326097

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research, 17,* 793–817. doi:10.1016/S0926-6410(03)00203-9

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491–1494. doi:10.1037/0278-7393.27.6.1491

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1,* 69–84. doi:10.1007/s11409-006-6894-z

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276.

Dijksterhuis, A. (2007). When to sleep on it. *Harvard Business Review, 85,* 30–32.

Ebbesen, E. B., & Rienick, C. B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology, 83,* 745–762. doi:10.1037/0021-9010.83.5.745

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review, 16,* 88–92. doi:10.3758/PBR.16.1.88

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition, 38,* 951–961. doi:10.3758/MC.38.7.951

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 288–299. doi:10.1037/0096-1523.1.3.288

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528. doi:10.1037/0033-295X.98.4.506

Glisky, E. L., Schacter, D. L., & Tulving, E. (1986). Learning and retention of computer related vocabulary in memory-impaired patients: Method of vanishing cues. *Journal of Clinical and Experimental Neuropsychology, 8,* 292–312. doi:10.1080/01688638608401320

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107,* 311–327. doi:10.1037/0033-2909.107.3.311

Hoffrage, U., & Pohl, R. (2003). Research on hindsight bias: A rich past, a productive present, and a challenging future, *Memory, 11,* 329–335. doi:10.1080/09658210344000080

Hollingworth, H. L. (1913). Experimental studies in judgment [Special issue]. *Archives of Psychology, 4*(Whole No. 29).

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modulates the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558. doi:10.1080/09541440601056620

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. doi:10.1037/0096-3445.126.4.349

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51,* 481–537. doi:10.1146/annurev.psych.51.1.481

Kornell, N., & Metcalfe, J. (2006). "Blockers" do not block recall in tip-of-the-tongue states. *Metacognition and Learning, 1,* 248–261. doi:10.1007/s11409-007-9003-z

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47,* 211–232.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279–308. doi:10.1007/BF01320096

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68,* 522–528. doi:10.1037/0022-0663.68.5.522

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. doi:10.1037/0033-295X.104.2.211

Lhyle, K. G., & Kulhavy, R. W. (1987). Feedback processing and error correction. *Journal of Educational Psychology, 79,* 320–322. doi:10.1037/0022-0663.79.3.320

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494–513. doi:10.1080/09541440701326154

McDaniel, M. A., & Fisher, R. P. (1991). Test and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201. doi:10.1016/0361-476X(91)90037-L

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131,* 349–363. doi:10.1037/0096-3445.131.3.349

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18,* 159–163. doi:10.1111/j.1467-8721.2009.01628.x

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132,* 530–542. doi:10.1037/0096-3445.132.4.530

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52,* 463–477. doi:10.1016/j.jml.2004.12.001

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14,* 225–229.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37,* 1077–1087. doi:10.3758/MC.37.8.1077

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology, 19,* 743–768. doi:10.1080/09541440701326063

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science, 32,* 1133–1147. doi:10.1080/03640210802353016

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

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. doi:10.1016/S0022-5371(80)90266-2

O'Neill, M. E., & Douglas, V. I. (1996). Rehearsal strategies and recall performance in boys with and without attention deficit hyperactivity disorder. *Journal of Pediatric Psychology, 21,* 73–88. doi:10.1093/jpepsy/21.1.73

Paller, K. A., Kutas, M., & Mayes, A. R. (1987). Neural correlates of encoding in an incidental learning paradigm. *Psychophysiology, 67,* 360–371.

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences, 6,* 93–102. doi:10.1016/S1364-6613(00)01845-3

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. doi:10.1037/0278-7393.31.1.3

Payne, D. G. (1987). Hypermnesia and reminiscence in recall: A historical and empirical review. *Psychological Bulletin, 101,* 5–27. doi:10.1037/0033-2909.101.1.5

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Sanna, L. J., & Schwartz, N. (2006). Metacognitive experiences and human judgments: The case of hindsight bias and its debiasing. *Current Directions in Psychological Science, 15,* 172–176. doi:10.1111/j.1467-8721.2006.00430.x

Schwartz, B. L. (2002). *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval.* Mahwah, NJ: Erlbaum.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 4,* 592–604.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87,* 116–124. doi:10.1037/h0030186

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19,* 645–647. doi:10.1111/j.1467-9280.2008.02136.x

Werth, L., & Strack, F. (2003). An inferential approach to the knew-it-all-along phenomenon. *Memory, 11,* 411–419. doi:10.1080/09658210244000586

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 345–353. doi:10.1037/0096-1523.4.2.345

Received May 9, 2010
Revision received October 4, 2010
Accepted October 11, 2010 ■