Understanding the Hypercorrection Effect: Why High-Confidence Errors are More Likely to be Corrected

by

Lisa K. Fazio

Department of Psychology and Neuroscience
Duke University

Date:_______________________
Approved:

___________________________
Elizabeth Marsh, Supervisor

___________________________
David Rubin

___________________________
Ian Dobbins

___________________________
Gavan Fitzsimons

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Psychology and Neuroscience in the Graduate School
of Duke University

2009

ABSTRACT

# Understanding the Hypercorrection Effect: Why High-Confidence Errors are More Likely to be Corrected

by

Lisa K. Fazio

Department of Psychology and Neuroscience
Duke University

Date:_______________________
Approved:

___________________________
Elizabeth Marsh, Supervisor

___________________________
David Rubin

___________________________
Ian Dobbins

___________________________
Gavan Fitzsimons

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor
of Philosophy in the Department of
Psychology and Neuroscience in the Graduate School
of Duke University

2009

# Abstract

The hypercorrection effect refers to the finding that high-confidence errors are more likely to be corrected after feedback than are low-confidence errors (Butterfield & Metcalfe, 2001). In 5 experiments I examine the hypercorrection effect, offer possible explanations for why the effect occurs, and examine the durability of the effect. In Experiment 1, I replicated the hypercorrection effect and showed that delaying the feedback does not reduce the effect. In a secondary item analysis I also showed that the effect is not caused by "tricky" questions. In Experiments 2 and 3, I showed that subjects are more likely to remember the source of the feedback after both high-confidence errors and low-confidence correct responses. Subjects' memory was most accurate when there was a discrepancy between their expectation and the actual feedback. This pattern of results suggests that subjects pay more attention to surprising feedback and that this leads to the hypercorrection effect. In Experiment 4, I showed that the hypercorrection effect also occurs for episodic false memories, showing that the effect is not limited to general knowledge questions. Finally, in Experiment 5, I examined the durability of the effect. Initial high-confidence errors that are corrected after feedback remain corrected one week later. Overall, the hypercorrection effect is robust. The effect occurs with multiple types of materials and its effects persist over a one week delay. Furthermore, the hypercorrection effect is caused by subjects paying more attention to feedback that is unexpected.

# Contents

## List of Tables

# List of Figures

# Acknowledgements

First and foremost I would like to thank my advisor Elizabeth Marsh. She has been everything that one could want in a mentor and more. Since we first worked together my junior year of college, Beth has been a constant source of advice and encouragement. I would not be the researcher that I am now without her guidance along the way. I am very proud to be her first graduating student and I know that there will be many more successful students in the years to come.

I would like to thank all of my committee members, past and present, David Rubin, Ian Dobbins, Gavan Fitzsimons and Patricia Bauer for all of their helpful comments and critiques of this research. I am also in debt to Peter Ornstein and Amy Needham for their consistent help and support as I expanded my research to developmental issues.

I have known since high school that I would grow up to research something and that passion for research came directly from my parents. It was because of their example that I entered graduate school, and it was thanks to their love and support that I finished it. I would also like to thank my brothers for being a constant reminder that life consists of more than academia. I am so proud of you both. To Paul, thank you for keeping me sane during this past year. No matter how crazy things got, you always knew that everything would turn out okay.

My thanks go out to Roddy Roediger, James Wertsch and Elizabeth Marsh for helping me to decide what it was that I wanted to research. It was their class on Trauma and Memory that started my fascination with how our memories can be changed and altered. I would also like to thank Kathleen McDermott for helping to further my interest in memory research while I was an undergraduate.

To my girls, Adrianna, Sarah and Lin, I would not have survived graduate school without your constant support and friendship.  We must plan another girls' night soon.  My thanks to all of my current and past lab mates, Mike Hudson, Holli Sink, Barbie Huelser, Andrea Eslick, Aaron Johnson, Anna Goswick, Sharda Umanath and Andrew Butler.  All of you have contributed to this dissertation in one way or another.  Finally, I would like to thank the American Psychological Association for their financial support of this research.

# 1. Introduction

As we go through our daily life, it is easy to forget how often we use our general world knowledge and how much new knowledge we learn every day. For example, stopping at the bagel shop in the morning I know that I'll get 13 bagels if I order a baker's dozen. When planning for a winter trip to Boston, I know that it will probably snow and that I should bring a coat and hat. Listening to the evening news, I know that Iran is a country in the Middle East with nuclear ambitions and I might learn that its current president is Mahmoud Ahmadinejad. Yet, people do not have perfect knowledge about the world around them. As we go about our lives and interact with the world, we discover errors in our knowledge that we have to correct.

This false information can come from many different sources; one's friends, a novel, one's own faulty memory. Even news programs, which strive to be accurate, can contain false information. For example, in the days following the invasion of Iraq in 2003 many events were reported in the news only to be retracted shortly thereafter. Lewandowsky and colleagues looked at whether people could identify true and fictional events that occurred shortly after the invasion (Lewandowsky, Stritzke, Oberauer, & Morales, 2005). An example of a true event would be "A 19-year-old female U.S. Prisoner of War was rescued from an Iraqi hospital by Special Forces and flown out of Iraq for medical treatment," while a false event would be "Iraqi troops poisoned a water supply station before withdrawing from the outskirts of Baghdad." They found that, in general, people were able to distinguish between the true and fictional events showing that they had learned facts from the news. More interesting was their examination of events that had been reported as true and later retracted. These included statements such as "During the first few days of the war, an entire Iraqi division (some 8,000 soldiers) was captured and/or surrendered to the allies." Participants tested in

America were likely to rate these events as true, even when they remembered that they had been retracted. Interestingly, participants from Germany, who were more suspicious of the war and events surrounding it, correctly rated the retracted items as false. This experiment eloquently shows how difficult it can be to remove incorrect information once it has been encoded.

Given that errors can be persistent, what is the best way to correct errors in the knowledge base? One simple and effective way to correct errors is by providing feedback to the learner. There are many ways of presenting feedback, however, and some are more effective than others. In the sections that follow I summarize the history of feedback research and then discuss which methods of feedback are most effective.

## *1.1 History of Feedback Research*

### 1.1.1 Early Research

The first researcher to examine the effects of feedback on learning was Edward Thorndike. In his research on animals, Thorndike discovered that the association between a stimulus and response was learned faster when the animal obtained a reward following the response. These findings lead to Thorndike's Law of Effect: stimulus-response connections that are accompanied by or followed by a "satisfying state of affairs" will be strengthened and connections accompanied by an annoying state of affairs will be weakened (Thorndike, 1911). Thorndike quickly applied his laws of learning to educational research and his ideas were extremely influential in the educational community (Thorndike, 1913). He believed that by giving feedback would both reinforce correct responses and help to weed out incorrect responses.

The major consequence of Thorndike's research was that researchers began to examine ways of giving students feedback immediately after being tested. One of these methods was to use testing machines. Developed by Sidney Pressey, these machines allowed students to continue to answer a question until they got the answer correct (Pressey, 1926, 1927). In addition, the machines could be set up to release candy as a reward for answering the question correctly. Thus, the machine complied with two of Thorndike's laws of learning. The last response was always correct (Law of Recency) and correct responses were rewarded (Law of Effect). Pressey believed that his machines would revolutionize education, however, educators quickly switched to cheaper methods for giving immediate feedback and his machines never caught on. The idea of answer-until-correct feedback, however, was very influential and educators began to use punchboards as an inexpensive way to provide immediate feedback (e. g. Angell & Troyer, 1948; Jensen, 1949; Morgan & Morgan, 1935).

### 1.1.2 Feedback as Reinforcement

By the 1950s the idea that feedback's main purpose was as reinforcement had become the dominant view. Gone was Thorndike's original idea that feedback would serve to reinforce correct responses and help to eliminate errors. Instead, feedback was believed to function solely as a reinforcer for correct responding. Given that rewards are most effective when they occur immediately following the desired behavior, giving immediate feedback was thought to be critical(Renner, 1964). Researchers believed that feedback would be ineffective unless it was received within seconds of the student answering the question (Skinner, 1954). The typical time lags observed in the school, where a student might not receive feedback for minutes or even days was thought to be very ineffective. Talking about the current state of the schools, Skinner writes, "Many seconds or minutes intervene between the child's response and the teacher's reinforcement . . . It is surprising that this system has any effect whatsoever" (Skinner, 1954, p. 91).

In order to solve this problem, Skinner invented his own teaching machine (Skinner, 1958). Skinner's machines differed in a number of ways from those proposed by Pressey. While Pressey's machine was focused on testing students after they had learned the material, Skinner focused on the learning process itself. Using a method called programmed instruction, students would be presented with short frames of information (normally only one or two sentences long). After each frame was a question that was designed to be very easy to answer. Students would then check to see that they got the answer correct and move onto the next frame. An example programmed learning text on high-school physics is shown below in Table 1.

**Table 1: Sample programmed instruction text from Skinner (1958).**

| Sentence to be completed | Word to be supplied |
|---|---|
| The important parts of a flashlight are the battery and the bulb. When we "turn on" a flashlight, we close a switch which connects the battery with the _____. | bulb |
| When we turn on a flashlight, an electric current flows through the fine wire in the ____ and causes it to grow hot. | bulb |
| When the hot wire glows brightly, we say that it gives off or sends out heat and ____. | light |
| The fine wire in the bulb is called a filament. The bulb "lights up" when the filament is heated by the passage of a(n) _____ current. | electric |

By making the questions simple and repetitive, researchers were able to virtually eliminate errors, which were thought to be aversive effects. Rather, the learners' behavior was to be shaped by operant conditioning. By reinforcing each of the small steps students were expected to learn more effectively. These programmed learning texts became very popular and were used in many school districts during the 1960s (Benjamin, 1988).

### 1.1.3 Feedback as Information

Beginning in the sixties, however, researchers began to question whether operant conditioning could properly explain students' behavior. One large problem was that researchers found that programmed learning texts that were presented without feedback were just as effective as texts presented with feedback (e. g. Feldhausen & Birt, 1962; Moore & Smith, 1961; Ripple, 1963). In addition, while the conditioning literature suggests that feedback should have its greatest effects as a reinforcer following correct answers, feedback has been repeatedly shown to have greater effects on incorrect answers than on answers that are initially correct (Anderson, Kulhavy, & Andre, 1971; Guthrie, 1971; Pashler, Cepeda, Wixted, & Roher, 2005). Finally, researchers began noticing that delaying the presentation of feedback did not impair performance and in some cases it was actually more beneficial than immediate feedback (Atkinson, 1969; Brackbill & Kappy, 1962; Kulhavey & Anderson, 1972). This finding went against the behaviouralist idea that the feedback needed to occur immediately following a response in order to function as a reinforcer.

These findings led a shift in the literature to begin thinking of feedback as information. Feedback is thought to confer two things, first knowledge of whether a response is correct or incorrect and second, knowledge of the correct answer. It is no longer thought of as having a role as a motivator or reinforcer. As such, feedback is thought to have its greatest effect on correcting errors, while having a small, but noticeable, effect on retaining correct responses. With this shift in thinking, researchers began to focus more on the interaction between the learner, the type of test and the type of feedback. Below I review three of the main variables that are thought to influence the effectiveness of feedback.

## *1.2 Type of Feedback*

When deciding what type of feedback to present to a learner, one aspect that must be considered is the content of the feedback. Students can simply be told that their answer is correct or incorrect (right/wrong feedback), they can be told what the correct answer is (answer feedback), or they can be told what the correct answer is and why that answer is correct (expanded feedback). Many studies have shown that providing answer feedback is beneficial in correcting errors (for reviews see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavey & Stock, 1989; Mory, 2004). Below, I examine more closely the benefits of right/wrong feedback and expanded feedback.

### 1.2.1 Right/Wrong Feedback vs. No Feedback

The first question is whether receiving right/wrong feedback can help students to correct their errors. This question is important because teachers often use right/wrong feedback rather than answer feedback when grading homework, exams and quizzes. It takes much less time to mark an answer as correct or incorrect than to write out the correct answer. While right/wrong feedback clearly provides less information than answer feedback, it does contain some useful information. The question is whether simply knowing if one's response is correct or incorrect can improve later performance.

There is evidence that right/wrong feedback does have some benefits. For example, low-confidence correct answers are more likely to persist onto a second test following right/wrong feedback as compared to a no feedback condition (Fazio, Huelser, Johnson, & Marsh, 2010). In terms of error correction, however, right/wrong feedback is often no more effective than receiving no feedback. In one prototypical experiment, Pashler and colleagues had subjects learn foreign language word pairs. Subjects then took two initial tests with answer feedback, right/wrong feedback or no feedback. On a final test held one week later, subjects who received right/wrong feedback corrected no more errors than subjects who received no feedback (Pashler et al., 2005). We recently found similar null results with both word pairs and with factual material learned from short passages (Fazio et al., 2010). In fact, in a meta-analysis, Kluger and DeNisi (1996) found that correct answer feedback reliably improves performance on a second test while subjects receiving right/wrong feedback showed no improvement. There is, however, one study showing right/wrong feedback to be more beneficial than receiving no feedback. Roper (1977) tested subjects on definitions of statistical concepts and found that subjects corrected more errors following right/wrong feedback as compared to no feedback. As of now it is unclear why subjects in Roper's experiment benefited from right/wrong feedback when subjects learning a variety of other topics have shown no benefit.

Some situations have been identified, however, where receiving right/wrong feedback is better than receiving no feedback. One of these situations is when the tests are multiple-choice rather than cued recall. When the number of answer choices is constrained, then receiving right/wrong feedback can be informative. If a subject narrows the answer choices down to two options, then learning that one option is incorrect can help the subject to choose the correct answer on a second test (Hanna, 1976; Marsh, Lozito, Bjork, & Bjork, submitted). Importantly, however, answer feedback is always more effective than right/wrong feedback (Fazio et al., 2010; Marsh et al., submitted; Pashler et al., 2005; Roper, 1977).

### 1.2.2 Answer feedback vs. Expanded Feedback

The next question is how expanded feedback compares to answer feedback. One might assume that more information is always better, however, this assumption has not always held up experimentally. Overall, the research on expanded feedback is sparse and very inconsistent (Kulhavey & Stock, 1989). Expanded feedback has been shown to have negative effects (Kulhavey, White, Topp, Chan, & Adams, 1985), to be no more effective than answer feedback (Gilman, 1969) and to have positive effects (Moreno, 2004).

In terms of negative effects, Kulhavy and colleagues found that students who were told why each of the multiple-choice lures was incorrect actually answered fewer questions correctly on a second test than students who were only told the correct answer (Kulhavey et al., 1985). The researchers provide two possible explanations for their surprising result. The first is that the addition information about the incorrect answer may have simply made that incorrect option more memorable. The authors do not report, however, how many of the initial errors persist onto the later test. Thus, one cannot examine if errors are more likely to persist after the expanding feedback. The second explanation is that subjects simply have too much to focus on in the expanded feedback condition, thus they end up skimming the feedback instead of deeply encoding it. While subjects did spend more time on the expanded feedback than on the answer feedback, it is unclear whether they were truly processing the additional information.

While expanded feedback can sometimes be ineffective or harmful, there are also occasions when expanded feedback may be useful. One key is that additional information may be unnecessary if the second test contains the exact same questions as the initial test. If all the subject needs to learn is the correct answer then the extra information provided by expanded feedback is unnecessary. When the second test contains different questions, however, then expanded feedback may be more useful. Using a multimedia computer program about botany, Moreno (2004) found that expanded feedback was more useful than answer feedback. Students who were told the correct answer along with why their initial answer was incorrect performed better on a later transfer test than subjects who were only told the correct answer. While answer feedback is very useful for learning the correct answer to a specific question, expanded feedback in more effective in teaching the underlying concepts behind an answer.

The benefits of different types of feedback may also depend on the types of errors being made. Tatsuoka and colleagues have found that feedback has little to no effect on serious errors, but for non-serious errors providing feedback is effective in reducing the number of errors produced. The researchers had eighth graders solve addition problems using signed numbers (e.g. -10+ 8 = __). Errors that quickly diverged from the correct solution path were labeled as serious errors (e.g. -10 + 8 = -18), whereas errors in the non-serious group followed most of the steps correctly (e.g. -10 + 8 = 2). For these non-serious errors, they found that expanded feedback was slightly better than correct answer feedback and correct answer feedback was considerably better than right/wrong feedback (Birenbaum & Tatsuoka, 1987). In contrast, the serious errors remained no matter what type of feedback was given.

In addition to providing the learner with additional information, expanded feedback can be induced by having the learner actively process the feedback. Phye and Andre (1989) showed that feedback is more effective when subjects are forced to process the feedback. Subjects who used the feedback to manually correct their incorrect responses scored better on a second test than subjects who simply studied the feedback. If subjects are processing the feedback, however, then it is unclear whether additional, deeper processing is helpful. Working with general knowledge questions, McDaniel and Fisher (1991) found no benefit of having subjects give a reason why the feedback was true. Subjects who elaborated on the feedback did no better on a second test than subjects who simply repeated the feedback. Importantly, there was no measure of the quality of subjects' responses. Therefore, it is possible that the subjects were not taking the task seriously and generating helpful information.

In contrast, Lhyle and Kulhavy (1987) did find benefits of additional feedback processing. Subjects who had to unscramble the provided words in order to discover the feedback corrected more errors on a second test than subjects who copied the feedback in its unscrambled form. Importantly, the benefit only occurred when subjects were forced to write down the unscrambled sentence. When responding was covert there was no difference between the two feedback groups. Together these findings suggest that deeper encoding of the feedback does improve later retention, but only if subjects treat the encoding task seriously.

## *1.3 Timing of Feedback*

A second consideration is when the learner should receive the feedback. Two contrasting views exist in the literature. The first is that students should receive feedback as quickly as possible (ideally immediately following each answer). The second is that students benefit more from delayed feedback. A number of studies have examined this issue over the years with wildly inconsistent results (for a review see Kulik & Kulik, 1988).

As mentioned previously, initially feedback was discussed in terms of behaviorist ideas of reinforcement. As such, the expectation was that feedback needed to be immediate in order to serve as a proper reinforcer. Researchers began to notice, however, that delaying the feedback did not have detrimental effects (Atkinson, 1969; Brackbill & Kappy, 1962; Kulhavey & Anderson, 1972). In fact, performance was often improved following delayed feedback (e. g. Brackbill, Wagner, & Wilson, 1964; More, 1969; Sturges, 1972).

An alternate view began to develop that not only was delaying feedback not harmful, it actually improved learning. One of the most prominent researchers who emphasized the benefits of delaying feedback was Raymond Kulhavy. With his interference-perseveration hypothesis (IPH), Kulhavy proposed the only comprehensive model of why delayed feedback would be beneficial (Kulhavey & Anderson, 1972). According to IPH, delayed feedback is more effective than immediate feedback because subjects begin to forget their incorrect responses over the delay. Thus, when the feedback is presented, subjects in the delay condition suffer from less proactive interference than subjects in the immediate feedback condition. Supporting this conclusion, subjects correct more of their initial errors following delayed feedback than following immediate feedback.

While the interference-perseveration hypothesis does have some empirical support, there is also evidence that discounts the explanation. For example, Peeck and colleagues have shown that when tested after a delay, subjects can readily identify their incorrect responses on the initial test (Peeck, van den Bosch, & Kreupeling, 1985). In addition, the delayed feedback is actually more effective for questions where the subject can identify their initial error (Peeck & Tillema, 1978).

Despite the large number of studies showing that delayed feedback is more effective than immediate feedback, there are also a number of studies showing the reverse (Kulik & Kulik, 1988). One reason for the confusion is that there is no single definition of immediate or delayed feedback. In some experiments, immediate feedback occurs after each item and delayed feedback is given at the end of the test (Butler, Karpicke, & Roediger, 2007) or delayed feedback can even occur as soon as 10 seconds after the subject gives a response (Brackbill & Kappy, 1962). In other experiments, immediate feedback occurs directly following a test and delayed feedback is presented days later (White, 1968). In addition, the studies vary in materials, in the type of questions asked and in whether the students actually process the feedback or if a graded test is simply handed back and ignored. Therefore, the variety of results is not surprising. Overall, however, the research supports the idea that delayed feedback reliably produces results that are as good, if not better, than immediate feedback (Butler et al., 2007).

## *1.4 Learner's Current Knowledge*

A third factor that determines the effectiveness of feedback is the current knowledge of the learner. It is this factor that will be the focus of this dissertation. Feedback likely has differential effects on high-confidence and low-confidence errors. The correction processes necessary when a student is initially guessing are different than the processes necessary when their current knowledge is incorrect. Students' confidence in their answers is based on two factors. First, students rely on the amount of information that they are able to retrieve to support their answer (Koriat, Lichtenstein, & Fischhoff, 1980; Tsai, Klayman, & Hastie, 2008). Second, students are more confidence in responses that come to mind quickly (Kelley & Lindsay, 1993; Nelson & Narens, 1990). Thus, responses that are given with high confidence, those that are retrieved quickly and with a lot of support, should be very entrenched in memory. Many theories of memory, therefore, would suggest that these high-confidence errors would be the most difficult to later correct (i.e. McGeogh, 1942; Raaijmakers & Shiffrin, 1981). The argument is that errors made with high confidence are firmly established in our memories, and thus difficult to eradicate from our knowledge base.

Intriguingly, several studies have shown that high-confidence errors are actually more likely to be corrected after feedback than are low-confidence errors. In an early demonstration, participants read short paragraphs about the eye, then answered multiple-choice questions, rated their confidence in each answer, and received feedback about the correct answers (Kulhavey, Yekovick, & Dyer, 1976). On a final multiple-choice test that repeated the same 30 questions, participants corrected more of their high-confidence errors than their low-confidence errors. More recently, Butterfield and Metcalfe (2001) found the same effect with different stimuli. In their experiment, participants answered general world knowledge questions such as "What poison did Socrates take at his execution?" Participants rated their confidence in each response and then were told the correct answer to each question. Similar to Kulhavy et al., Butterfield and Metcalfe (2001) found that high-confidence errors were more likely to be corrected on a retest than were low-confidence errors. The authors named this finding the *Hypercorrection Effect*.

Why is it that these high-confidence errors, which should be firmly established in memory and difficult to update, are instead more likely to be corrected than are low-confidence errors? One possibility is that participants attend more to unexpected feedback, with positive consequences for memory. In other words, when a participant makes an error with high confidence, the feedback is surprising, leading the learner to more deeply encode the feedback. This surprise hypothesis is similar Kulhavy's model of how feedback affects learning (Kulhavey, 1977; Kulhavey, Yekovick et al., 1976), and owes a debt to Rescorla and Wagner's (1972) model of animal learning (which stated that learning occurs fastest when events violate the organism's expectations). Kulhavy proposed that a large discrepancy between the participant's initial beliefs and the correct answer leads the participant to expend more effort to correct the misunderstanding.

One prediction of this model is that participants should choose to spend more time studying the feedback after a high-confidence error; this was confirmed by Kulhavy (1977). However, the hypercorrection effect occurs even when the duration of the feedback is held constant (as in Butterfield & Metcalfe, 2001), and so the challenge is to find evidence for surprise when differential study times are not possible. Some support comes from neuroimaging data; for example, Butterfield and Mangels (2003) used ERPs to show that high-confidence errors elicited activity in frontal areas that have been linked to novelty in other studies (see Butterfield (2003) for a similar result using fMRI).

In the five experiments that follow I examine the hypercorrection effect in more detail. Experiment 1 provides an independent replication of the hypercorrection effect and examines the effect of delaying the feedback. I also reanalyze the data to show that the effect is not due to a subset of "tricky" questions. Experiments 2 and 3 show that unexpected feedback is better remembered, supporting the surprise hypothesis. Experiment 4 expands the hypercorrect effect to the domain of false memories and Experiment 5 examines the durability of the effect over a week's delay.

## 2. Experiment 1: Immediate vs. Delayed Feedback

To start, we were interested in replicating Butterfield and Metcalfe's findings. The finding was intriguing, but there were no independent replications of their results. Furthermore, we were concerned about the number of observations at each confidence level. Given their method, subjects could have answered as few as fifteen questions incorrectly on the initial test (Butterfield & Metcalfe, 2001). With 7 possible levels of confidence, each confidence bin could have had very few observations in it. In addition, as shown in Figure 1, in the original paper the hypercorrection effect was heavily reliant on the results for the highest confidence errors. If one removes the results for confidence level 7, then the line appears flat. Given that high-confidence errors are rare, this critical data point is likely based on very few answers.

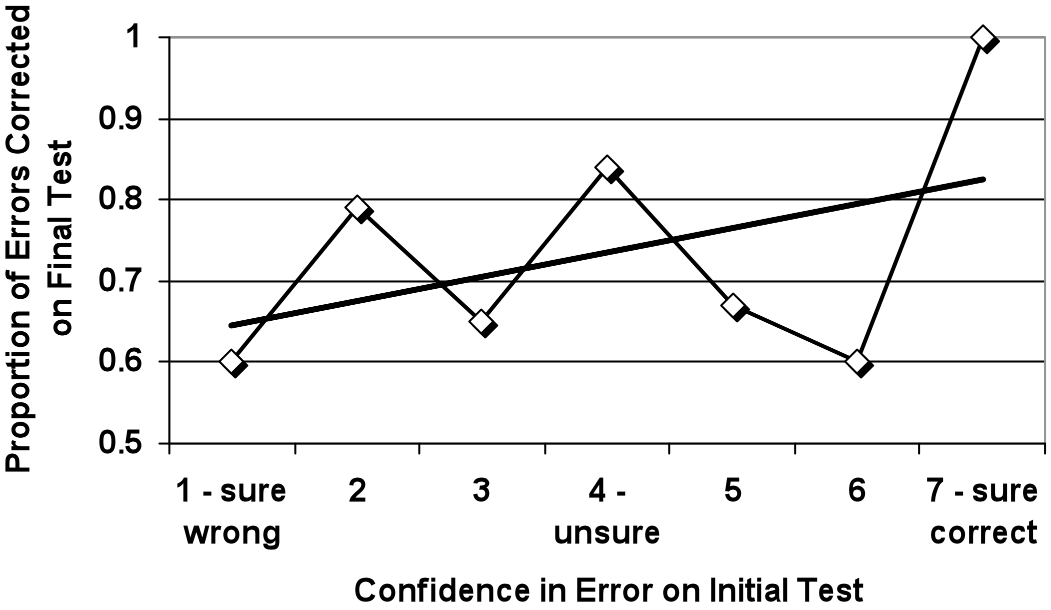


**Figure 1: Average proportion of the errors on the first test that were corrected on the second test for each confidence level. Bolded line is the best fitting trend line. (data from Butterfield & Metcalfe, 2001).**

For these reasons we were interested in both replicating the effect and exploring one possible cause of the effect.  As mentioned above, the surprise hypothesis suggests that the hypercorrection effect occurs because students are surprised to be told that they were incorrect following a high-confidence error.  They then pay more attention to the feedback, with benefits for later memory.  Therefore, one should be able to reduce the hypercorrection effect by reducing the students' surprise at being told that they were incorrect.

In order to reduce the contrast between the student's response and the correct answer we delayed the presentation of the feedback until the end of the test.  Subjects in the immediate feedback condition received feedback after each question while subjects in the delayed feedback condition did not receive feedback until the end of the test. After the delay, we predicted that the contrast between the subject's answer and the correct answer would be less noticeable and that the hypercorrection effect would be reduced.  In addition, we had subjects answer over 100 general knowledge questions so that we would have a wide distribution of responses across the different confidence levels.

## *2.1 Methods*

### 2.1.1 Participants

Forty-six Duke University undergraduates participated in the experiment for partial fulfillment of a course requirement.  Half of the participants received immediate feedback and the other half received delayed feedback.

### 2.1.2 Materials

One hundred forty general knowledge questions were selected from the Nelson and Narens (1980) norms.  The questions varied in difficulty; on average, 40% of participants in the norming study answered these items correctly (ranging across items from 0% to 92% correct).

### 2.1.3 Procedure

The experiment began with a general knowledge test. Participants were told they were to answer a series of questions and rate their confidence in each answer. They were warned that some of the questions would be difficult and that they should make educated guesses, or else respond "I don't know." Participants in the immediate feedback condition were told they would see the correct answer after each question and that they should try to remember the answers because they would be asked the same questions later in the experiment. Participants in the delayed feedback condition were not told about the feedback and the second test until after they answered all of the questions.

Each of the general knowledge questions appeared on the screen in a random order. Participants typed their response to each question and rated their confidence using a 7-point scale. Following Butterfield & Metcalfe (2001), the scale ranged from 1 (sure wrong) to 4 (unsure) to 7 (sure correct). For subjects in the immediate feedback condition, the correct answer appeared for 5 seconds after each confidence rating was recorded. This feedback took the form of a sentence, and was presented regardless of whether the question was answered correctly or not. For example, if the question was "What's the longest river in South America?" then the feedback was "Amazon is the longest river in South America." Subjects in the delayed feedback condition first answered all of the general knowledge questions, giving a confidence rating for each answer. After finishing the test the subjects were told that they would now see the correct answers for all of the questions that they had just answered. The feedback was presented immediately following the subject's completion of the initial test. As in the immediate feedback condition, the feedback was presented in the form of a sentence and remained on the screen for 5 seconds. The feedback sentences appeared in a random order.

Following the general knowledge test, all of the subjects solved visuo-spatial puzzles for 4 minutes as a short filler task, and then retook the general knowledge test. The final test was identical to the initial test except no feedback was provided. Participants were then debriefed and thanked for their participation.

## 2.2 Results

Unless otherwise noted, differences were significant at the .05 level.

Performance on the initial test did not differ across the two conditions. Participants correctly answered 43% of the initial questions in the immediate feedback condition and 46% in the delayed feedback condition, $t < 1$. Confidence on the initial test was also similar across the conditions, averaging g 4.25 in the immediate condition and 3.95 in the delayed condition, $t(44) = 1.13$, $SED = .27$, $p = .27$. Participants were well calibrated in their use of the confidence scale for both conditions; the average within-participant gamma correlation between initial test accuracy and confidence was .79 for the immediate feedback condition and .86 for the delayed feedback condition, $t(44) = 1.49$, $SED = .05$, $p = .14$.

Feedback improved performance similarly for both the immediate and delayed feedback conditions. Participants in the immediate feedback condition improved from 43% correct on the initial test to 83% correct on the final test, $t(22) = 26.76$, $SEM = .01$. In the delayed feedback condition, participants improved from 46% correct on the initial test to 82% correct on the final test, $t(22) = 15.79$, $SEM = .02$. Final test performance did not differ between the two conditions, $t < 1$.

Of primary interest was whether we replicated the hypercorrection effect and if the effect was reduced following delayed feedback. For each of the seven confidence levels on the first test, we examined the proportion of errors that were successfully corrected on the second test. As shown in Table 2, there were adequate numbers of incorrect answers at each confidence level for both conditions. [1]

**Table 2: Average number of incorrect answers given at each level of confidence on the initial test separated by feedback condition. (Experiment 1).**

| | Confidence | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1<br>sure wrong | 2 | 3 | 4<br>unsure | 5 | 6 | 7<br>sure correct |
| Immediate FB | 13.30 | 6.35 | 4.57 | 9.70 | 5.26 | 3.08 | 6.30 |
| Delayed FB | 7.61 | 6.43 | 3.52 | 9.30 | 3.87 | 2.83 | 2.70 |

Figure 2 shows hypercorrection for both conditions: participants corrected more of the errors that had been committed with high confidence than those made with low confidence. The mean within-subject gamma correlation between initial confidence and proportion of errors later corrected was significantly positive in both the immediate feedback, $\gamma = .22$, $t(22) = 2.59$, $SEM = .09$, and the delayed feedback conditions, $\gamma = .18$, $t(22) = 2.15$, $SEM = .08$. There was no difference between the gamma correlations in the two conditions, $t < 1$.

[1] We only counted erroneous responses as incorrect answers. "Don't know" responses were counted separately and not included in these analyses.

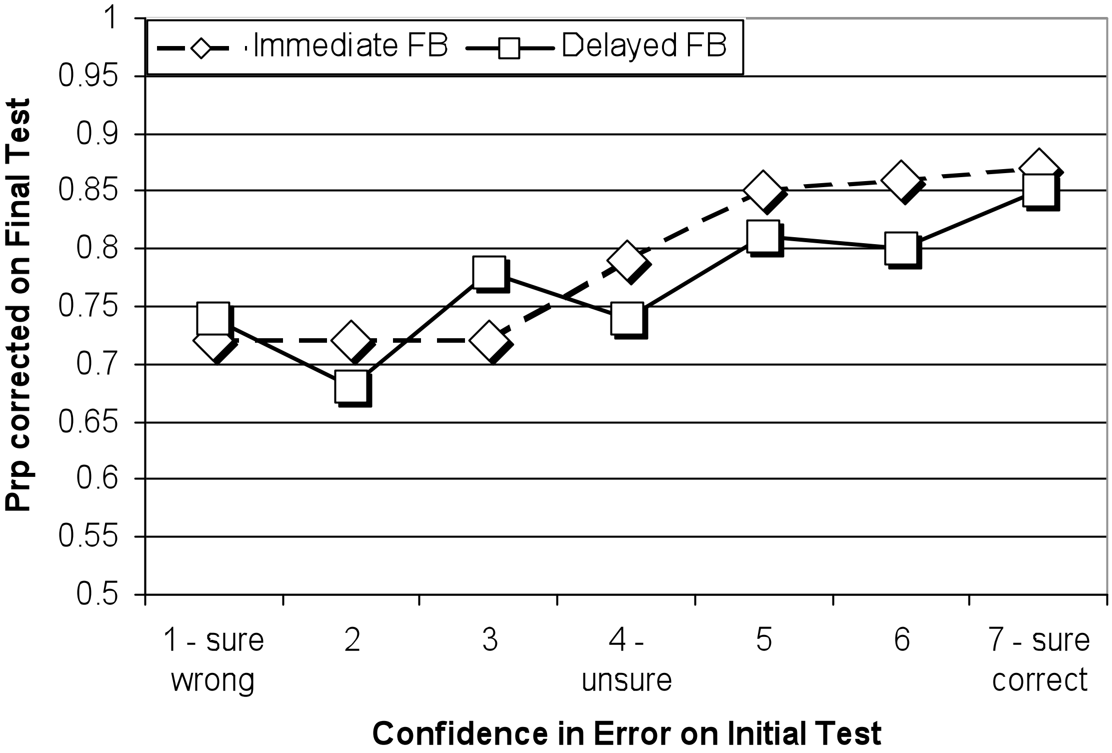


**Figure 2: Average proportion of the errors on the first test that were corrected on the second test for each confidence level for both the immediate and delayed feedback conditions. (Experiment 1)**

## 2.3 Discussion

We replicated the hypercorrection effect; high confidence errors were more likely to be corrected than low confidence errors. In addition, using 140 questions guaranteed that there were an adequate number of observations at each level of confidence. On average subjects answered 4.5 questions incorrectly with the highest level of confidence.

Contrary to our expectations, both the immediate and the delayed feedback groups showed strong evidence for the hypercorrection effect. Simply delaying the feedback by a few minutes was not enough to reduce the effect. Our initial assumption that delaying the feedback would reduce its impact may have been mistaken.  We expected that by the end of the test, the answers that the subjects had previously given would have begun to fade from memory and thus the feedback would be less surprising.  Contrary to our assumption, prior research has shown that subjects are able to identify their prior answers to test questions even days later (Peeck & Tillema, 1978). In addition, it has recently been shown that delayed feedback is as effective or even more effective at correcting errors than immediate feedback (Butler et al., 2007).   Using the same delay as in our experiment (the feedback was presented after each question or at the end of the test) Butler and colleagues found that subjects corrected more errors following delayed feedback.  Thus, our manipulation may have increased the feedback's effectiveness rather than decreasing its effect.  It now seems that delaying the feedback presentation is unlikely to reduce the subject's sense of surprise when told that their high-confidence response was incorrect.

## *2.4 Item Effect Analysis*

One criticism that has been levied against the hypercorrection effect is that it only occurs for "tricky" questions (Pashler, personal communication). These are questions where a very popular and common answer is actually wrong, e.g. "What's the capital of Australia?" (The correct answer is Canberra, not Sydney). It is possible that the hypercorrection effect only occurs for these "trick" questions that invite a high confidence error. For example, Pashler et al. (2005) only found a trend for the hypercorrection effect when subjects were tested and retested on foreign language word pairs.

### 2.4.1 Methods

We reexamined the data from Experiment 1 looking for questions where subjects were likely to make high confidence errors. Specifically we looked for questions where subjects were more likely to answer the question incorrectly with high confidence than with low confidence.  For each question we determined how many subjects answered it incorrectly with each level of confidence. We then calculated the correlation between confidence level and the number of incorrect answers. Forty-two questions had a positive correlation between the two variables - meaning that people were more likely to answer the question incorrectly with high confidence than with low confidence. We then redid the analysis from Experiment 1 eliminating these 42 "trick" questions.

### 2.4.2 Results

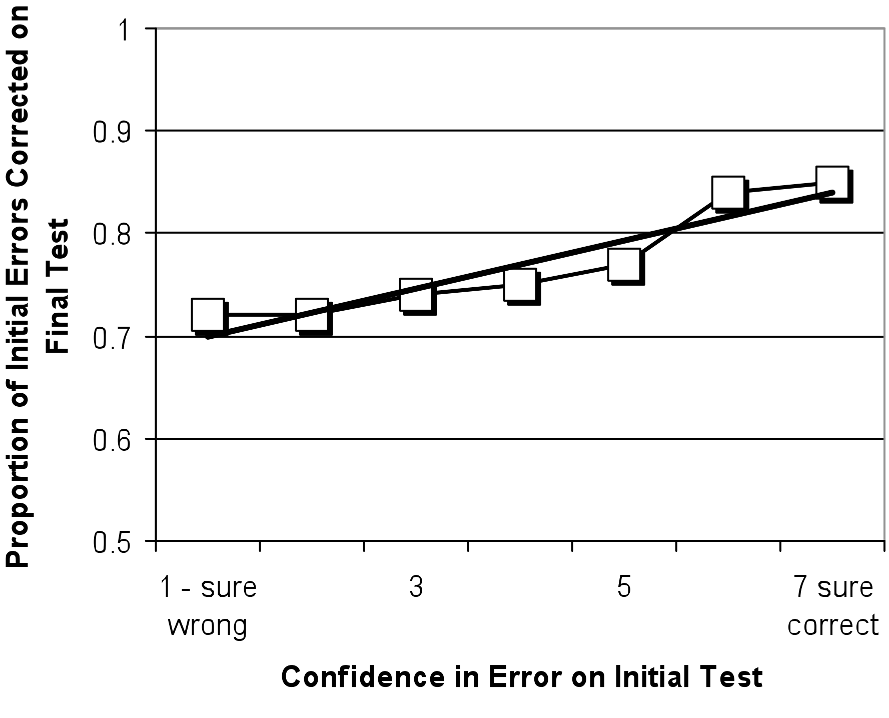


**Figure 3: Average proportion of the errors on the first test that were corrected on the second test for each confidence level. Bolded line is the best fitting trend line. (Experiment 1, select questions).**

As shown in Figure 3, the hypercorrection effect still occurred after eliminating the questions that reliably produced high confidence errors. The mean within-subject gamma correlation between initial confidence and proportion of errors later corrected was significantly positive, $\gamma = .18$, $t(45) = 2.41$, $SEM = .08$. In addition, when analyzed alone the gamma for the tricky items ($\gamma = .13$, $t < 1$) was no larger than the gamma for the non-tricky items.

### 2.4.3 Discussion

This analysis shows that the hypercorrection effect is not limited to "trick" questions. In fact, the trick questions were no more likely to produce a hypercorrection effect than the non-tricky questions. Rather, the hypercorrection effect occurs whenever a subject gives an incorrect answer with high confidence. It does not matter whether that question tends to universally elicit high confidence errors or if the subject's error is idiosyncratic. High confidence errors are more likely to be corrected for both types of questions.

# 3. Experiments 2 and 3: Remembering Feedback Sources

The next set of experiments examined more directly whether the surprise hypothesis could explain the hypercorrection effect. That is, are high confidence errors more likely to be corrected because of the subject's surprise at being told that his answer is wrong?

When answering questions, the feedback may be surprising in two different situations. In addition to high-confidence errors, an individual should also be surprised when he or she believes a response to be a guess (and rates confidence as low), and yet finds out that the guess was correct. A test-feedback-retest paradigm only allows examination of errors, as low-confidence correct answers do not need to be corrected. But the surprise hypothesis predicts that both situations should have consequences for attention, and in turn for later memory. Butterfield and Metcalfe (2006) used this logic in a pair of experiments in which participants did a tone detection task while answering the initial general knowledge questions, and then were retested with full attention. During the initial test, participants simply had to press a key whenever they heard a tone; critical was participants' ability to detect tones played concurrently with feedback. Surprising feedback was presumed to divert attention from the tone detection task. Consistent with this, participants missed more tones when the feedback revealed an error made with high confidence. In contrast, for correct answers, tone detection was better for high-confidence responses than for correct guesses. Overall, tone detection was negatively related to performance on the retest, suggesting that participants encoded the feedback at the expense of detecting the tone.

Our research also takes advantage of the fact that the surprise hypothesis predicts increased attention (and memory) for both high-confidence errors and low-confidence correct answers. However, instead of using distraction from another task to infer attention to the feedback, we chose a more direct measure of attention to the feedback, one that could be measured for both correct answers and errors: memory for the feedback's appearance. This dependent measure has been used in emotion research, with the result that memory is better for the surface features (e.g., font colors) of attention-grabbing emotional and taboo words than neutral words (Doerkenson & Shimamura, 2001; MacKay & Ahmetzanov, 2005). This is a measure of source memory or memory for the "conditions under which a memory is acquired" (Johnson, Hashtroudi, & Lindsay, 1993, p. 3). We are using a broad definition of source memory that includes everything that gets encoded about the feedback other than its content. Our argument is that source memory will be better when feedback is surprising, for both correct and incorrect answers.

In Experiment 2, participants answered general world knowledge questions, rated their confidence in each answer, and received feedback in the form of the correct answer to each question. Critically, the feedback appeared either in red or green font. After a short delay, participants identified whether each correct answer had been presented in green or red during the feedback phase. If a discrepancy between the participant's expectation and the feedback leads to a deeper encoding of the feedback, then source memory should be better for high-confidence errors *and* low-confidence correct responses, as compared to low-confidence errors or confident correct answers. In other words, when the feedback confirmed participants' beliefs, they should have paid less attention to it, resulting in lower memory for the feedback's appearance. This same relationship was expected in Experiment 3, where male and female voices delivered the feedback. One group of participants completed a source test (as in Experiment 2): for each correct answer, they identified whether it had been spoken in a male or female voice. Other participants were simply retested on the general knowledge questions. Thus Experiment 3 was designed to generalize the relationship between confidence and appearance memory to a different source judgment, as well as to demonstrate the standard hypercorrection effect in our modified paradigm. Because of the similarities between the two experiments, they will be discussed together in a single general discussion.

## *3.1 Experiment 2*

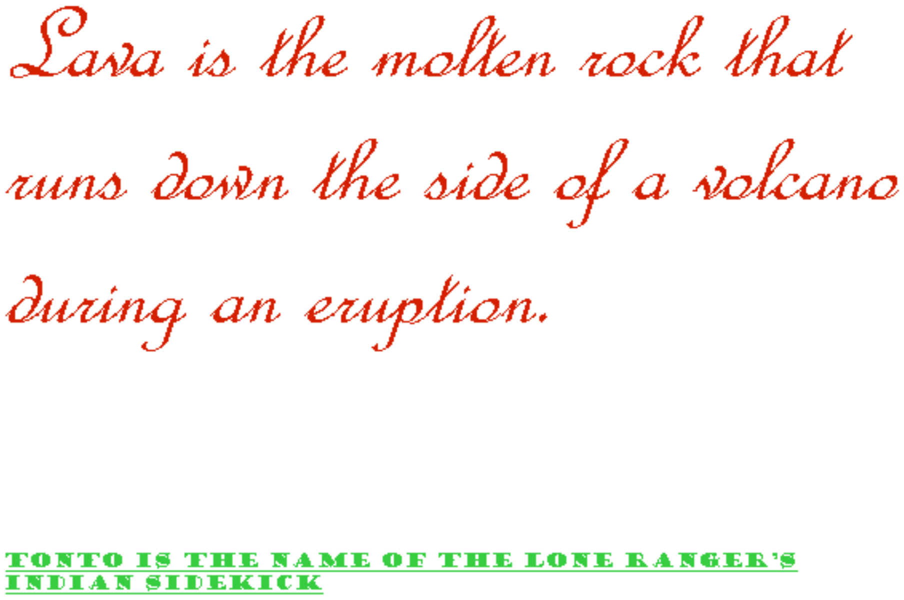
### 3.1.1 Method

#### 3.1.1.1 Participants

Forty-six Duke University undergraduates participated in the experiment for partial fulfillment of a course requirement. Seventeen additional participants were tested but performed at chance on the source discrimination task; thus, their data was excluded from the analyses. Chance was defined as answering less than 55% of the source questions correctly. None of the participants was color-blind.

#### 3.1.1.2 Materials

We again used the one hundred forty general knowledge questions used in Experiment 1. The feedback appeared in Times New Roman font. It was either red and italicized in 64 pt. font or green, underlined and bolded in 12 pt. font. Samples of both the red and green feedback are shown in Figure 4. All other text was presented in light blue 24 pt. font.



Lava is the molten rock that
runs down the side of a volcano
during an eruption.

TONTO IS THE NAME OF THE LONE RANGER'S
INDIAN SIDEKICK

**Figure 4: Examples of the red and green feedback presented in Experiment 2.**

### 3.1.1.3 Procedure

The experiment began with a general knowledge test. Participants were told they were to answer a series of questions and rate their confidence in each answer. They were warned that some of the questions would be difficult and that they should make educated guesses, or else respond "I don't know." Furthermore, they were told they would receive feedback on their answers and that they would later take a second test. Critically, the nature of the second test was never mentioned.

Participants typed their response to each question, and rated their confidence using a 7-point scale. Following Butterfield & Metcalfe (2001), the scale ranged from 1 (sure wrong) to 4 (unsure) to 7 (sure correct). The correct answer appeared for 5 seconds after each confidence rating was recorded. This feedback took the form of a sentence, and was presented regardless of whether the question was answered correctly or not. For example, if the question was "What's the longest river in South America?" then the feedback was "Amazon is the longest river in South America." For half of the items, the feedback was presented in the red font, whereas for the other half feedback appeared in the green font.

Immediately following the general knowledge test, participants completed a source test on their memory for the feedback's appearance. The feedback sentences were tested one at a time, in random order, in the light blue font. For each item, participants identified whether the feedback had been presented previously in red or green font. After the source test, participants were debriefed and thanked for their participation.

## 3.1.2 Results

### 3.1.2.1 Initial Test

On the initial test, participants answered an average of 43% of the questions correctly, and their average confidence was 4.11. Participants were well calibrated in their use of the confidence scale; the average within-participant gamma correlation between initial test accuracy and confidence was .78.

### 3.1.2.2 Source Test

Participants correctly identified the prior color of the feedback for 69% of the facts.

Of primary interest was the relationship between confidence on the initial test and performance on the source test. The surprise hypothesis predicts a different relationship between confidence and source memory for items answered correctly vs.

incorrectly on the initial general knowledge test. For general knowledge questions answered *correctly*, the feedback would have been unexpected for guesses, thus predicting better source memory for low-confidence correct answers than for high-confidence ones. In contrast, for general knowledge questions answered *incorrectly,* the feedback would have been surprising for high-confidence errors, thus predicting better source memory for high-confidence errors than low-confidence ones. In short, the surprise hypothesis predicts a negative relationship between source memory and confidence for items answered correctly on the initial general knowledge test, but a positive relationship for errors.

Figure 5 shows the relationship between source memory and confidence as a function of correctness on the initial general knowledge test. As predicted, source memory was highest when participants' confidence was mismatched with the accuracy of their original responses. For correct answers, lower confidence on the initial test was associated with better source memory. The mean within-subject gamma correlation between initial confidence and later source memory was significantly negative, $\gamma$ = -.19, $t$(45) = 2.61, *SEM* = .07. For incorrect answers, higher confidence was associated with better source memory, $\gamma$ = .12, $t$(45) = 2.23, *SEM* = .05.

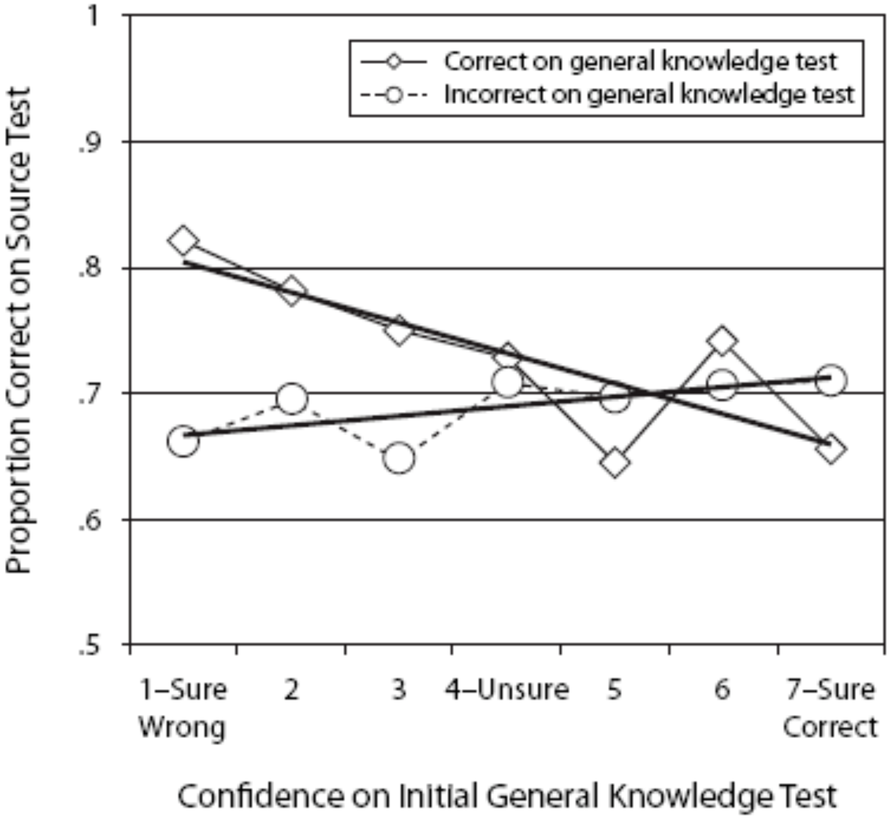


**Figure 5: Average proportion correct on the source test for each confidence level, as a function of whether the answer on the initial test was correct or incorrect. Bolded lines are the best fitting trend lines. (Experiment 2).**

A series of additional analyses were conducted to ensure that the key results were not due to differential memory for the red font (*M* = .72), which turned out to be more memorable than the green font (*M* = .65), *t*(45) = 3.07, *SEM* = .02. The reader will remember that half of the feedback statements were presented in red and half in green; but because we could not predict a priori an individual's responses nor their confidence in these responses, we could not counterbalance the font across the 14 cells. Critically, for errors, red and green feedback were not unequally distributed across the seven levels of confidence, *F*(6, 270) = 1.05, *Mse* = 2.65, *p* > .3, and thus better memory for red feedback could not explain the positive relationship between confidence and source memory found for errors. For correct responses, disproportionately more red feedback occurred in the high-confidence cells, *F*(6, 270) = 6.95, *Mse* = 2.42, but this is not concerning as the result predicted (and obtained) was in the opposite direction.

In short, the results were consistent with the surprise hypothesis: the relationship between confidence and source memory was positive for errors but negative for correct answers.

Because Experiment 2 focused on source memory, there was no measure of error correction. That is, a second general knowledge test was not administered after the source test, as the source memory test effectively presented the feedback for a second time. In Experiment 3, one group of participants took the source test and another group was retested on the general knowledge questions, to ensure that the changed paradigm did not eliminate the basic hypercorrection effect. In addition, the sources were made more distinctive in Experiment 3, to minimize loss of participants due to chance performance on the source test.

## *3.2 Experiment 3*

### 3.2.1 Method

#### 3.2.1.1 Participants

Seventy-two undergraduates participated in the experiment for partial fulfillment of a course requirement. In the final test phase, fifty participants took the source test. Six additional participants were tested in this condition but were excluded because they performed at chance on the source test (chance was defined as in Experiment 2). Twenty-two participants were in the retest condition; one additional participant was tested but excluded because he corrected all of his initial errors on the second test, making it impossible to calculate the relationship between his confidence in the initial error and the probability of it being corrected on the retest.

### 3.2.1.2 Materials

We used 120 of the original 140 questions from Experiments 1 and 2. We reduced the number of items to improve subject's memory for the feedback sources. The feedback was presented in one of two ways. For half of the items, a female voice read the feedback aloud, while a woman's picture appeared on the left side of the computer screen and the feedback printed in pink lettering appeared on the right. For the other half of the items, the voice was male and the computer screen showed a man's picture on the right and the feedback in blue lettering on the left. Examples of the male and female feedback are presented in Figure 6.

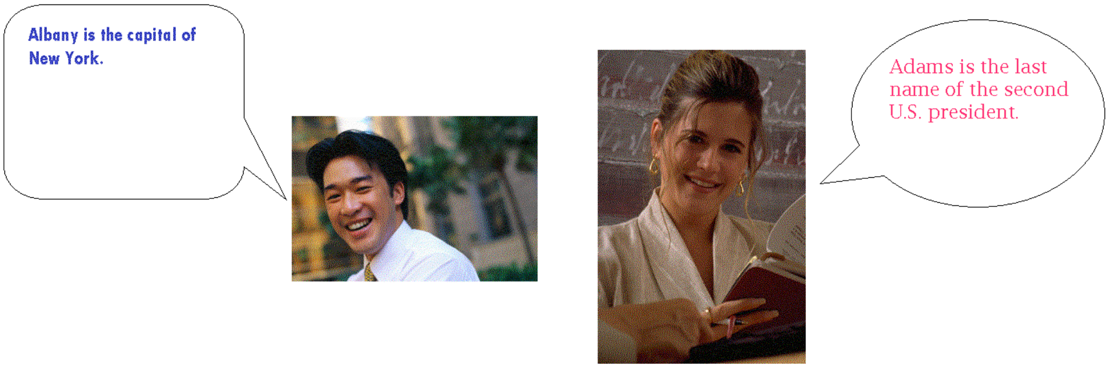

Albany is the capital of New York.
Adams is the last name of the second U.S. president.

**Figure 6: Examples of the male and female feedback presented in Experiment 3.**

### 3.2.1.3 Procedure

As in Experiment 2, participants answered a series of general knowledge questions, rated their confidence in each response, and then received feedback. To improve source memory, the feedback appeared for 6 seconds instead of 5 seconds as in Experiment 2.

After the general knowledge test, participants in the source memory condition immediately began the source test. The feedback sentences appeared on the screen in a neutral font and the participants identified whether the male or the female source had presented the feedback.

Participants in the retest condition solved visuo-spatial puzzles for 4 minutes before taking their final test, as pilot testing showed that participants were at ceiling without a short filler task. These participants then retook the general knowledge test, which was identical to the first test except no feedback was provided.

## 3.2.2 Results

### 3.2.2.1 Initial Test

Performance on the initial test did not differ across the two conditions. Participants correctly answered 42% of the initial questions in the source condition and 43% in the retest condition, $t < 1$. Confidence on the initial test was also similar across the conditions, averaging 4.01 in the source condition and 4.21 in the general knowledge retest condition, $t <1$. These values are similar to what was observed in Experiment 1, as were participants' confidence-accuracy correlations. The average within-subject gamma correlation between proportion correct and confidence on the initial test was .81 in the source condition and .76 in the retest condition, $t(70) = 1.35$, *SEM* = .04, $p$ =.18.

### 3.2.2.2 General Knowledge Retest

For the participants in the general knowledge retest condition, we compared performance on the initial test to performance on the final test. Feedback improved performance, with participants answering 80% of the questions correctly on the second test as compared to 43% on the initial test, $t(21) = 26.17$, *SEM* = .01.

Of primary interest was whether the hypercorrection effect occurred. For each of the seven confidence levels on the first test, we examined the proportion of errors that were successfully corrected on the second test. Figure 7 shows hypercorrection: participants corrected more of the errors that had been committed with high confidence than those made with low confidence. The mean within-subject gamma correlation between initial confidence and proportion of errors later corrected was significantly positive, $\gamma = .23$, $t(21) = 2.27$, $SEM = .10$.

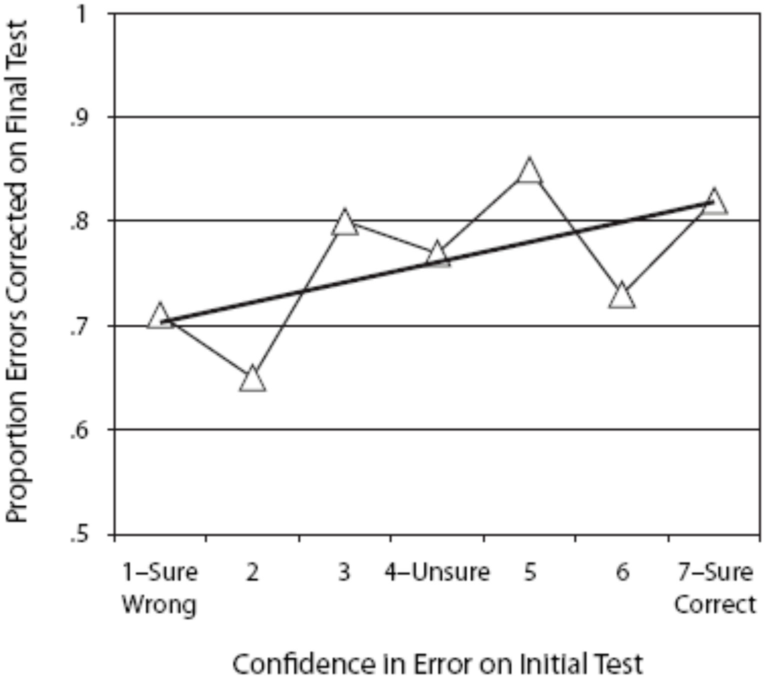


**Figure 7: Average proportion of the errors on the first test that were corrected on the second test for each confidence level. Bolded line is the best fitting trend line. (Experiment 3).**

### 3.2.2.3 Source Test

For the participants in the source condition, we examined memory for the source of the feedback. On average, participants correctly identified the source for 68% of the facts.

As in Experiment 2, of primary interest was the relationship between confidence on the initial test and later memory for the source of the feedback. Replicating Experiment 2, there was a negative relationship between confidence and source memory for correct answers and a positive relationship between confidence and source memory for errors, as shown in Figure 8. After answering a question correctly, participants were more likely to remember the source of the feedback if they had answered with low confidence than if they had answered with high confidence $\gamma$ = -.28, *t*(49) = 4.24, *SEM* = .07. The pattern was opposite for errors; participants were more likely to remember the source of the feedback if they had answered with high confidence than if they had answered with low confidence $\gamma$ = .12, *t*(49) = 2.18, *SEM* = .06.

As in Experiment 2, we conducted additional analyses to ensure that our results were not due to one source being more memorable than the other. We found that the male source was more likely to be accurately identified (*M* = .70) than the female source (*M* = .66), *t*(49) = 2.53, *SEM* = .02. However, the male and female feedback was not unequally distributed across the confidence levels for correct answers, *F*(6, 294) = 1.80, *Mse* = 2.29, $p > .1$ or for errors, $F < 1$. Thus, better memory for the male source cannot explain our results.

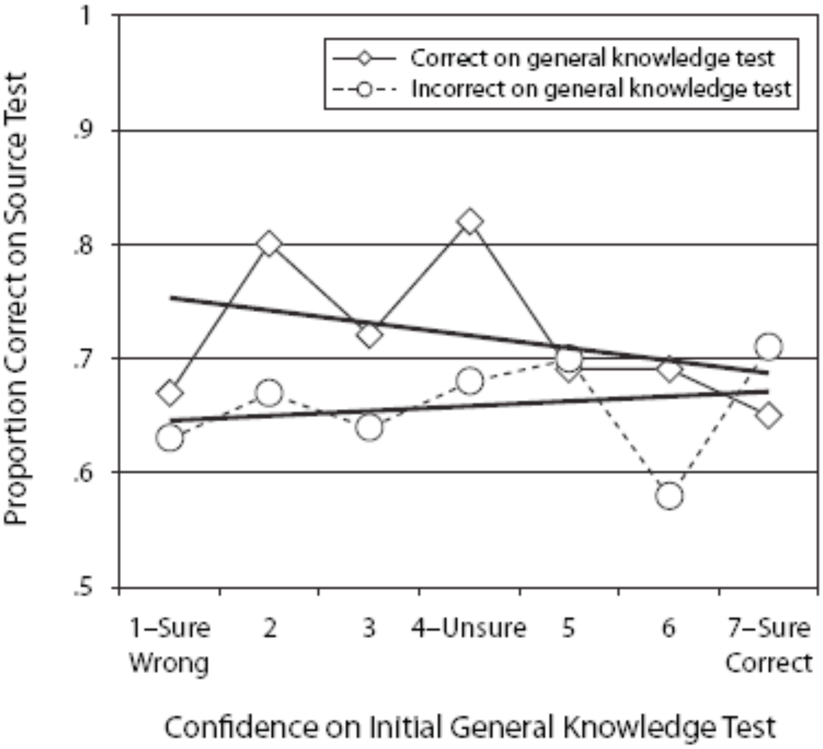


**Figure 8: Average proportion correct on the source test for each confidence level split by whether the answer on the initial test was correct or incorrect. Bolded lines are the best fitting trend lines. (Experiment 3).**

### 3.3 Discussion

In two experiments, surprising feedback improved memory for both the surface features and the content of presented feedback. In Experiment 2, participants were better able to remember the color of feedback when it was incongruent with their expectations. That is, source memory was better for feedback that had been presented in response to correct guesses or errors made with high confidence. In Experiment 3, participants showed improved memory for both the content and the source of the feedback. Participants were more likely to correct high-confidence errors than low-confidence errors and they were more likely to remember the source of the feedback when it was unexpected.

While the observed relationships between initial confidence and source memory were relatively small, they were as predicted in both experiments and occurred for both correct and erroneous answers. It is not surprising that the effects were smaller ones given that remembering the appearance of the feedback was not participants' main task. The participants in both experiments were lead to believe that they would be retested on the general knowledge questions - the source memory test was unexpected. Thus, most of the participants' additional attention should have been, and was, directed towards the content of the surprising feedback, rather than its surface features. This can be seen most clearly in Experiment 3 where memory for the content of the feedback (the correct answer) increased more than 10% across the confidence levels, while source memory increased less than 5%.

These experiments support the surprise hypothesis, which states that unexpected feedback leads to a greater expenditure of effort to encode that feedback, with positive consequences for memory. Data across laboratories are converging in support of the surprise hypothesis. Putting these results together, a consistent picture is emerging: feedback can be surprising (Butterfield & Mangels, 2003), leading to a focus on the feedback (the present studies) at the expense of other tasks (Butterfield & Metcalfe, 2006).

In addition to the surprise hypothesis, there is at least one other possible explanation of the hypercorrection effect. The knowledge hypothesis posits that confidence tends to be correlated with how much a participant knows generally about the target domain (Butterfield & Metcalfe, 2001). The argument is that if participants have little knowledge about a domain, then they have nothing with which to associate the incoming information. In other words, it will be more difficult to integrate the correct answer into their semantic memory if it is an unfamiliar domain. Although our experiments were not designed to test the knowledge hypothesis, it is not immediately clear what the knowledge hypothesis would predict about memory for the source of the feedback. In particular we doubt that the knowledge hypothesis would predict a negative relationship between source memory and confidence in correct answers. Of course our data do not rule out the knowledge hypothesis, as the two hypotheses are not mutually exclusive. It is quite plausible that knowledge updating requires both deep encoding of the feedback and a knowledge structure that allows the new information to be easily assimilated – but our data suggest that differences in domain knowledge are unlikely to be solely responsible for the hypercorrection effect.

# 4. Experiment 4: Correcting False Memories

The previous two experiments provided support for the surprise hypothesis, but they did not explicitly address a second explanation of the hypercorrection effect, the knowledge hypothesis. As just described, the key assumption underlying the knowledge hypothesis is that high-confidence errors are more likely to occur when subjects are knowledgeable about the target domain. A subject is more likely to realize she is guessing when faced with *"What is the capital of Seychelles?"* than when answering *"What is the capital of Australia?"*, assuming she is more familiar with the country Australia than with Seychelles. Because the subject knows some facts about Australia, she can associate the feedback *"Canberra"* to her existing knowledge base. In contrast, she has no other information about Seychelles to associate to the feedback "*Victoria*" and thus will be less likely to retrieve this feedback later. In order to establish if prior knowledge was essential for the hypercorrection effect, we changed our materials. Instead of having subjects answer general knowledge questions where there is a natural correlation between subjects' confidence and their prior knowledge, we had subjects remember sentences. Thus, we switched from a task that referenced subjects' semantic memory to one that required episodic memory. While the subjects must reference their prior knowledge to understand the sentences, there is not the same distribution of prior knowledge across confidence levels as observed in the domain of semantic memory. Therefore, the knowledge hypothesis does not predict the correction of episodic memories to vary as a function of confidence.

In order to ensure that subjects incorrectly remembered some of the sentences with high confidence we used sentences that implied a non-stated action. For example, the sentence *"The karate champion hit the cinder block*" is often misremembered as *"The karate champion broke the cinder block"* (Brewer, 1977). Not only are these errors easily created, they often become vivid false memories held with high confidence (Brewer, Sampaio, & Barlow, 2005). These materials then gave us the added benefit of looking at whether subjects can correct high-confidence false memories such as the ones produced by these sentences.

Overall, false memories can be strikingly persistent. Warning subjects that their memories may be incorrect is rarely effective (McDermott & Roediger, 1998), especially when the warning occurs after the study phase (Greene, Flynn, & Loftus, 1982). Even re-exposure to the original events fails to eliminate false memories. A second chance to hear "*bed, rest, tired, awake…*" reduces but does not eliminate false memories for "*sleep*" (McDermott, 1996; J. M. Watson, McDermott, & Balota, 2004). Similarly, re-reading a prose passage does little to eliminate any errors made in initial free recall of the passage (e. g. Fritz, Morris, Bjork, Gelman, & Wickens, 2000). One problem may be that subjects fail to detect contradictions between their memories and the re-presented events. However, even when the errors are explicitly marked, many still persist to later tests. McConnell and Hunt (2007) had students score their recall of related words such as "*bed, rest, tired…*" by checking off correct responses and putting an X next to intrusions. Participants corrected 65% of their errors on a test two days later, but a third of the original intrusions re-appeared.  Despite the evidence that false memories are very difficult to correct, the hypercorrection effect suggests that false memories (which typically occur with high-confidence) would be corrected more often than other low-confidence memory errors.

We tested these predictions in an experiment exploring whether false memories were more likely to be corrected than lower confidence errors. Participants studied 48 sentences, each of which contained a pragmatic implication (e.g. "*The absent-minded professor didn't have his car keys*"). They then took a cued recall test on the sentences, rated their confidence in each answer, and read the correct answer. Of interest were participants' responses on a second cued recall test. To preview, we observed hypercorrection of false memories on the final test: false memories were more likely to be corrected than were low-confidence memory errors.

## *4.1 Method*

### 4.1.1 Participants

Forty-six Duke University undergraduates participated in the experiment for partial fulfillment of a course requirement. Subjects were tested individually or in groups of up to three people.

### 4.1.2 Materials

We used forty-eight sentences containing pragmatic implications (e.g., "*The karate champion hit the cinder block*"), each of which was paired with a cued recall prompt (e.g., *"The karate champion* ____*the cinder block"*). The materials came from McDermott and Chan (2006), whose subjects completed an average of 47% of the cued recall prompts with pragmatic implications rather than with studied words (across items, 21 to 91% were completed with pragmatic implications). On average, subjects needed to produce 2.08 words to complete the sentence.

### 4.1.3 Procedure

During the **study phase**, each sentence was presented for 3500 ms, with a 500 ms inter-stimulus interval. Subjects were told to read and remember the sentences for a later test.

After the study phase, subjects took the **initial cued recall test**. Each sentence fragment was tested individually and subjects were asked to complete each with the exact wording previously studied. Confidence in each answer was rated on a 7-point scale (1 = sure wrong, 4 = unsure, 7 = sure correct). The feedback then appeared for 4 seconds. The original sentence was re-presented with the previously missing portion bolded (e.g., "*The karate champion **hit** the cinder block*"). This process was repeated for each of the 48 sentences.

Subjects then participated in an unrelated experiment for 10 minutes before the **final cued recall test**. Each of the sentence fragments was tested in the same manner as on the initial test; subjects tried to complete each with the studied wording and rated their confidence in each answer. No feedback was presented during the final test. Subjects were then thanked and debriefed.

## 4.2 Results

Following McDermott and Chan (2006), each cued recall answer was categorized as correct, a pragmatic implication, or another error. Critically, pragmatic implications were defined a priori. Consider the sentence "*The karate champion ___ the cinder block*." A response was classified as a pragmatic implication if it stated that the block had been broken. For example, the verbs "*broke*", "*smashed*" and "*split*" were all scored as pragmatic implications. A response was scored as correct if it matched the original wording exactly ("*hit*"), varied slightly in form (e.g., tense) or differed only in auxiliary words (e.g. "*hit at*"). Finally, responses such as "*kicked*" were scored as other errors as they matched neither the correct answer nor the pragmatic implication. Two coders scored the data and agreed on 99% of the items; the author resolved the discrepancies.

As shown in Table 3, participants frequently completed the initial cued recall prompts with pragmatic implications. In fact, pragmatic implications (*M* = .51) were more common than correct answers (*M* = .25), *t*(45) = 8.23, *SEM* = .03, or other errors (*M* = .25), *t*(45) = 11.74, *SEM* = .02. However, participants successfully used the feedback to correct many errors. On the final test, participants answered more questions correctly (M

= .74) than with pragmatic implications (M = .14), *t*(45) = 17.24, *SEM* = .03. The feedback also helped participants to reduce other errors (*M* = .12), as compared to what was observed on the initial test (*M* = .25), *t*(45) = 9.27, *SEM* = .01. Pragmatic implications and other errors were directly compared in a 2 (error: pragmatic, other) x 2 (time of test: initial, final) ANOVA. While both types of errors dropped from the initial to the final test, *F*(1, 45) = 616.30, *MSE* = .01, $\eta_p^2$ = .93, this decrease was larger for pragmatic implications, as reflected in a significant interaction between error type and time of test, *F*(1, 45) = 136.58, *MSE* = .01, $\eta_p^2$ = .75.

**Table 3: Proportion of cued recall questions answered correctly, with a pragmatic implication, or with other errors. (Experiment 4).**

| | Cued Recall Test | |
|---|---|---|
| | Initial | Final |
| Correct | .25 | .74 |
| Pragmatic | .51 | .14 |
| Other error | .25 | .12 |

Overall, correct answers on the initial test were made with higher confidence (*M* = 3.91) than were the pragmatic implications (*M* = 3.72), *t*(45) = 3.59, *SEM* = .05, which in turn were made with higher confidence than other errors (*M* = 3.54), *t*(45) = 5.02, *SEM* = .04. Table 4 shows the distribution of the three responses across the seven confidence levels on the initial test. Subjects were more likely to answer correctly with high than low confidence, showing that they accurately used the confidence scale. Correspondingly, the mean within-subject gamma correlation between the number of correct responses and the confidence level was significantly positive, $\gamma$ = .42, *t*(45) = 7.38, *SEM* = .06.

**Table 4: Average number of correct, pragmatic and other wrong answers given at each level of confidence on the initial cued recall test. (Experiment 4).**

| | Confidence | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1<br>sure wrong | 2 | 3 | 4<br>unsure | 5 | 6 | 7<br>sure correct |
| Correct | 0.54 | 0.52 | 0.83 | 2.22 | 2.30 | 2.33 | 3.04 |
| Pragmatic | 4.00 | 2.72 | 2.98 | 6.59 | 3.98 | 2.57 | 1.46 |
| Other Error | 3.48 | 1.50 | 1.20 | 2.85 | 1.61 | 0.98 | 0.33 |
| | | | | | | | |
| Total | 8.02 | 4.74 | 5.01 | 11.66 | 7.89 | 5.88 | 4.83 |

Subjects were also calibrated when judging their confidence in pragmatic inferences and other errors. They made more errors with low confidence than errors with high confidence, yielding negative within-subject gamma correlations for both pragmatic responses, $\gamma = -.15$, $t(45) = 2.48$, $SEM = .06$, and other errors, $\gamma = -.33$, $t(45) = 5.86$, $SEM = .06$. However, subjects were less well calibrated for pragmatic implications than for other errors, as shown by the stronger negative correlation for other errors ($\gamma = -.33$) than for pragmatic responses ($\gamma = -.15$), $t(45) = 3.78$, $SEM = .05$. The smaller gamma for pragmatic implications supports that these items yielded more high-confidence errors (false memories).

Our main focus was on which errors were corrected on the final test, and whether there was hypercorrection of false memories. As shown in Figure 9, hypercorrection was observed; pragmatic implications made with high confidence (the false memories) were more likely to be corrected on the final test than were lower confidence memory errors. The within-subject gamma correlation between initial confidence and later correction was positive and significant, $\gamma$ = .13, *t*(44) = 2.05, *SEM* = .06. There were not enough other errors to examine the correlation between correction and confidence for these items; however, a similar result was obtained when the pragmatic implications were combined with other errors: the within-subject gamma correlation between initial confidence and later correction of errors was positive, $\gamma$ = .14, *t*(45) = 2.51, *SEM* = .06.

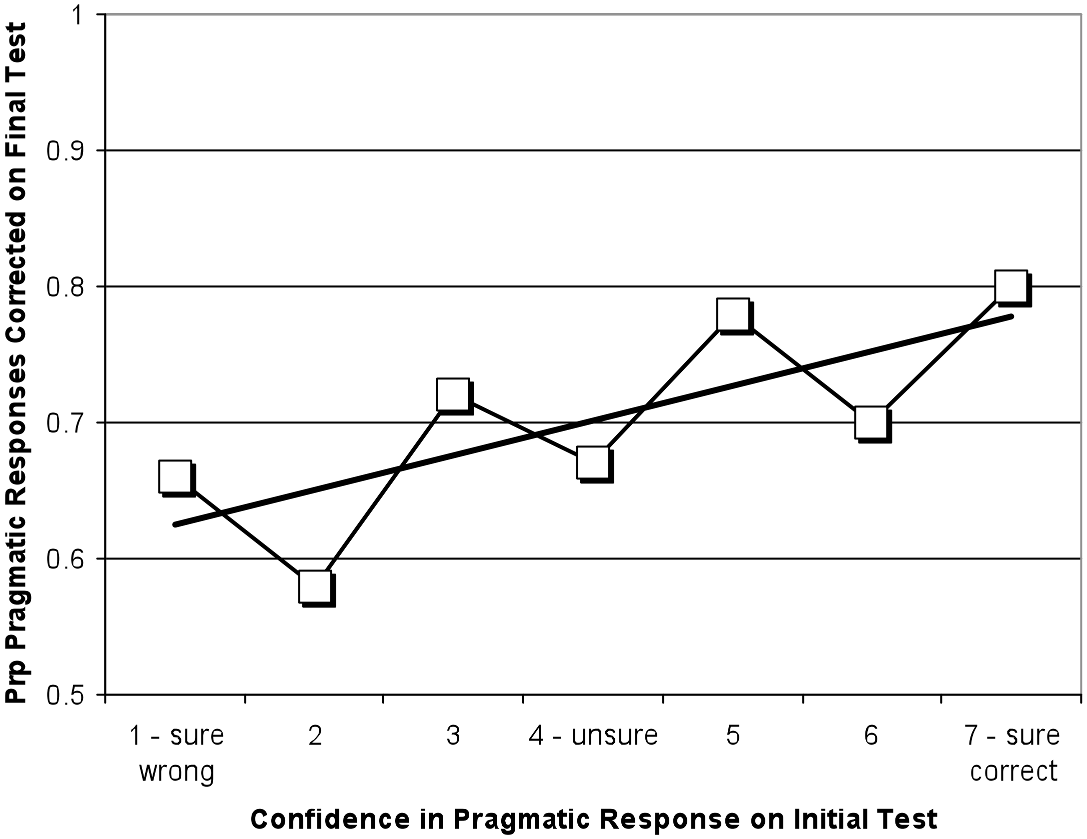


**Figure 9: Average proportion of the pragmatic responses on the initial test that were corrected on the final test, separated by initial confidence rating. Bolded line is the best fitting trend line. (Experiment 4).**

## 4.3 Discussion

With feedback, participants corrected more than two-thirds of their errors. Our success in correcting false memories is likely due to the targeted nature of the feedback, with the exact to-be-remembered wording appearing immediately after each response. The critical wording was bolded and easily compared to one's response; in contrast, subjects may fail to notice their errors when they re-read a passage or are re-exposed to a list (e. g. Fritz et al., 2000; McDermott, 1996). Correcting false memories requires directed feedback (as in the present study and those of McConnell & Hunt, 2007) or combining re-exposure with other accommodations, such as slowed presentation (e.g., Watson, McDermott, & Balota, 2004).

Most important was the finding of hypercorrection of false memories: after receiving feedback, subjects were more likely to correct false memories (made with high confidence) than erroneous guesses. This finding of hypercorrection for episodic memories means that differences in domain knowledge cannot be solely responsible for the hypercorrection effect. To be clear, false memory creation depends upon prior knowledge; a sentence like "*the karate champion hit the cinder block*" yields the inference "*the karate champion* ***broke*** *the cinder block*" because of stored knowledge about what typically happens when a block is hit. However, what is crucial for present purposes is that this prior knowledge is constant across the seven confidence levels. In other words, the lower correction rate for guesses cannot be due to an inability to associate the feedback to existing knowledge. To be clear, although this work suggests that domain knowledge cannot be the *only* factor driving hypercorrection, it does not imply that differences in domain knowledge *never* contribute to the effect.

Our finding has implications for correcting other episodic memory errors; corrections will be most likely when the feedback contradicts subjects' expectations. One interesting question is whether the result will hold for updating, more generally, even if the update is not correct. For example, consider the eyewitness who remembers a red car with high confidence; what happens when another witness insists it was a blue car? We know that witnesses reject blatantly contradictory misinformation (Loftus, 1979), but the hypercorrection effect suggests that confident memories may be more open to updating. Key is that the eyewitness must trust the other witness's account more than her own, perhaps because the other witness had a better view of the accident. If so, the hypercorrection effect predicts that the confident eyewitness would be more likely to switch her report to a blue car than if she had only guessed that the car was red.

# 5. Experiment 5: The Persistence of Corrected Errors

The previous four experiments have shown that the hypercorrection effect is a reliable effect and that it occurs for errors in both episodic and semantic memory. The next question is how effective the hypercorrection effect functions as a long-term learning mechanism. After a delay, does the original error return or does the correction persist? If the original error quickly returns, then the correction is only transient and does not cause any long-term changes to the learner's knowledge base. If, however, the correction remains, then the feedback did successfully alter the learner's knowledge. Furthermore, it suggests that high-confidence errors can be successfully eliminated through a simple feedback procedure.

There is prior research supporting both the prediction that the correction will persist and that the original error will reappear. The corrective effects of feedback are often seen even after a delay, suggesting that feedback can induce lasting change in a learner's knowledge base. Many studies have shown beneficial effects of feedback after a one-week delay (e.g. Butler et al., 2007; Butler & Roediger, 2008; Pashler et al., 2005). Unfortunately, the only experiments to examine delays of more than one week were designed to compare different types of feedback and thus do not contain a control group that was tested but received no feedback (Little, 1934; White, 1968). Therefore, it is impossible to determine if the gains in performance were due to the feedback or other learning processes. Nevertheless, there is clear evidence that corrective effects of feedback persist for at least one week.

In contrast, the subject's original high-confidence errors should be older memories that are very well established and thus would be likely to reappear following a delay. According to Jost's law, if two memories are of equal strength then the older memory will decay at a slower rate than the younger memory (Wixted, 2004). Thus, after a delay an older memory may be more likely to be retrieved than a younger memory. With the hypercorrection effect, the high-confidence original error is likely to be both older and more strongly encoded than the newly learned correct answer. Thus, the original error may be likely to reappear on the delayed test. In addition to Jost's law, work on spontaneous recovery also suggests that the original errors may reappear. Pavlov's early experiments with dogs showed that a conditioned response will often reappear, even after it appears to be extinguished (Pavlov, 1927). In addition, the initial response is more likely to reappear as the delay between the extinction trials and the second test increases (Rescorla, 2004). Clinical research also shows that while people initially respond very well to extinction treatments of phobias, their fears are likely to return at a later date (Lang, Craske, & Bjork, 1999). This research suggests that while the subjects may be able to correct their errors on an initial test, over time the original error (which is older and more firmly established in memory) will return.

There are only three studies have examined the effects of delay on the hypercorrection effect. In the first, Kulhavey and colleagues had subjects read short passages about the human eye and then answer multiple-choice questions on each passage (Kulhavey, Yekovich, & Dyer, 1976). Subjects rated their confidence in each answer and then received feedback about the correct answer. After the initial test, subjects were immediately retested on the same items. One week later, the subjects returned to the lab and answered the test questions for a third time. Overall, the researchers found that initial errors made with high confidence were more likely to be corrected than errors made with low confidence, both on the immediate retest and on the third test one week later. That is, they found a hypercorrection effect even after the delay. One confound in the experiment, however, is that subjects were able to study the feedback as long as they wanted. Furthermore, feedback that occurred following a high-confidence error was studied for longer than feedback following a low-confidence error. Therefore, the hypercorrection effect found on both the immediate and delayed tests could be due to the differential study times rather than differences in how the feedback was processed.

Two more recent studies on the effects of delay on the hypercorrection effect show conflicting results (Butler, Karpicke, & Roediger, 2008; Butterfield & Mangels, 2003). Unlike the Kulhavy experiment, where subjects took three tests (two in the immediate session and one after a week delay), these experiments involved only two tests. The initial test and feedback occurred in the first session and the second test occurred one week later. Butterfield and Mangels (2003) still found that initial high-confidence errors were more likely to be corrected than low-confidence errors, even after the one week delay. In contrast, Butler and colleagues (2008) did not find evidence for the hypercorrection effect with a one week delay in between the feedback and the second test. The researchers acknowledge, however, that they did not have very many observations at each confidence level. Thus, the effects of delay on the hypercorrection effect are still unclear.

In this experiment we were interested in a different question than the one asked by Butler et al (2008) and Butterfield and Mangels (2003). Instead of examining if the hypercorrection effect occurs with a delay between the feedback and the retest, we were interested in what happened to previously corrected errors following a delay. Thus, we used a method similar to the one used by Kulhavy (1976). We used a constant presentation rate for the feedback, however, rather than letting the subjects control the presentation rate. Subjects came into the lab and answered a series of general knowledge questions. After each question, subjects rated their confidence in their answer and received feedback about the correct answer. Following a five-minute filler task the subjects were retested on all of the questions. The subjects then left the lab and returned one week later to take a third test on all of the questions. We were interested in what would happen to the initial high-confidence errors that were corrected on the second test. Would these initial errors reappear on the third test, or would subjects continue to respond with the correct answer?

## *5.1 Method*

### 5.1.1 Participants

Thirty-five Duke University undergraduates participated in the experiment and received monetary compensation. One subject was eliminated because he or she only used two points on the 7-point confidence scale. Thus, the analyses below include thirty-four subjects.

### 5.1.2 Materials

Two hundred and twenty-five general knowledge questions were selected from a variety of sources including trivia books, published experiments and Internet searches (Abrams, Trunk, & Margolin, 2007; Berger, Hall, & Bahrick, 1999; Dahlgren, 1998; Nelson & Narens, 1980; Preston & Preston, 2005; Vigliocco, Vinson, Martin, & Garrett, 1999; Yaniv & Meyer, 1987). The questions touched on a wide variety of topics including geography, history, famous people, animals and sports. This experiment used more questions than Experiments 1-4. This increase ensured that we had sufficient numbers of correct and incorrect answers in each confidence bin, even after the delay.

### 5.1.3 Procedure

The experiment consisted of two sessions, separated by one week. The first session lasted approximately one and a half hours and the second session lasted around thirty minutes. In the first session, subjects answered each of the 225 general knowledge questions, e.g. "*What country is Copenhagen the capital of?*" Subjects were instructed that some of the questions would be very difficult, but that they should try to answer each question even if they had to guess. If they were unable to come up with an answer than they were told to write, "I don't know." After answering each question, they rated their confidence in that answer on a 7-point scale (1 = sure wrong, 4 = unsure, 7 = sure correct). The correct answer was then presented for three seconds, e.g. "*Copenhagen is the capital of Denmark*". Subjects were told that they should try to remember the feedback because they would be tested on the same questions later in the experiment.

After answering all of the questions, subjects solved visio-spatial puzzles for 4.5 minutes as a short filler task. They were then retested on all 225 cued recall questions. Subjects again rated their confidence in each answer, but no feedback was presented.

One week later, subjects returned for the second session of the experiment. They again answered all 225 cued recall questions and rated their confidence in each answer. No feedback was presented. After finishing the final test, the subjects were debriefed and thanked for their participation.

## 5.2 Results

To start, we conducted a one-way ANOVA on the proportion of questions answered correctly on each of the three tests. As shown in Table 5, subjects answered more questions correctly after the feedback was presented, $F(2, 66) = 752.58$, $MSE = .002$, $\eta_p^2 = .96$. Forty-six percent of the questions were answered correctly on the initial test and this increased to 85% on the second test, $t(33) = 31.39$, $SEM = .01$. Furthermore, the feedback was still effective one week later. On the final test subjects answered 77% of the questions correctly. This was both an increase from the initial test, $t(33) = 26.77$, $SEM = .01$, and a decrease from the second test, $t(33) = 10.45$, $SEM = .01$.

Following feedback, subjects also made fewer errors, as shown by the significant one-way ANOVA on incorrect responses across the three tests, *F*(2, 66) = 144.78, *MSE* = .003, $\eta_p^2$ = .81. Whereas the subjects answered 26% of the questions incorrectly on the initial test, only 4% of the questions were answered incorrectly on the second test, *t*(33) = 12.83, *SEM* = .02. There was a small increase in incorrect responses from the second test (*M* = .04) to the third test (*M* = .07), *t*(33) = 6.11, *SEM* = .005, but the proportion of incorrect answers on the third test remained far below the error rate on the initial test, *t*(33) = 11.53, *SEM* = .02.

**Table 5: Proportion of correct, incorrect and don't know responses on the three tests. Standard deviations are in parenthesis (Experiment 5).**

| | Initial Test | Second Test | Final Test |
|---|---|---|---|
| Correct | .46 (.13) | .85 (.08) | .77 (.10) |
| Incorrect | .26 (.12) | .04 (.04) | .07 (.05) |
| Don’t Know | .29 (.15) | .11 (.07) | .16 (.08) |

Following feedback, 83% of subjects’ initial errors were corrected on the second test. However, our main focus was if the errors corrected on the second test remained corrected on the third test. Overall, 84% of the initial errors that were corrected on the second test remained correct on the third test. In contrast, only a few initial errors reappeared on the third test after being corrected on the second test. Only 5% of corrected errors switched back to subjects’ initial error on the final test, far fewer than the 84% that remained corrected, t(33) = 36.20, SEM = .02. But some errors did reappear; the 5% was a significant increase from zero, t(33) = 6.53, SEM = .01. Overall, however, the vast majority of the corrected errors remained correct even after a week delay.

### 5.2.1 Results split by initial confidence

The next question involves looking specifically at the initial high-confidence errors. Were high-confidence errors more likely to be corrected than low-confidence errors on the second test? And did that pattern persist onto the third test? We first examine the hypercorrection effect from the initial test to the second test. As shown in Figure 10, high-confidence errors on the initial test were more likely to be corrected on the second test than were low-confidence errors. Correspondingly, the mean within-subject gamma correlation was significantly positive, $\gamma$ = .14, *t*(33) = 2.09, *SEM* = .07.

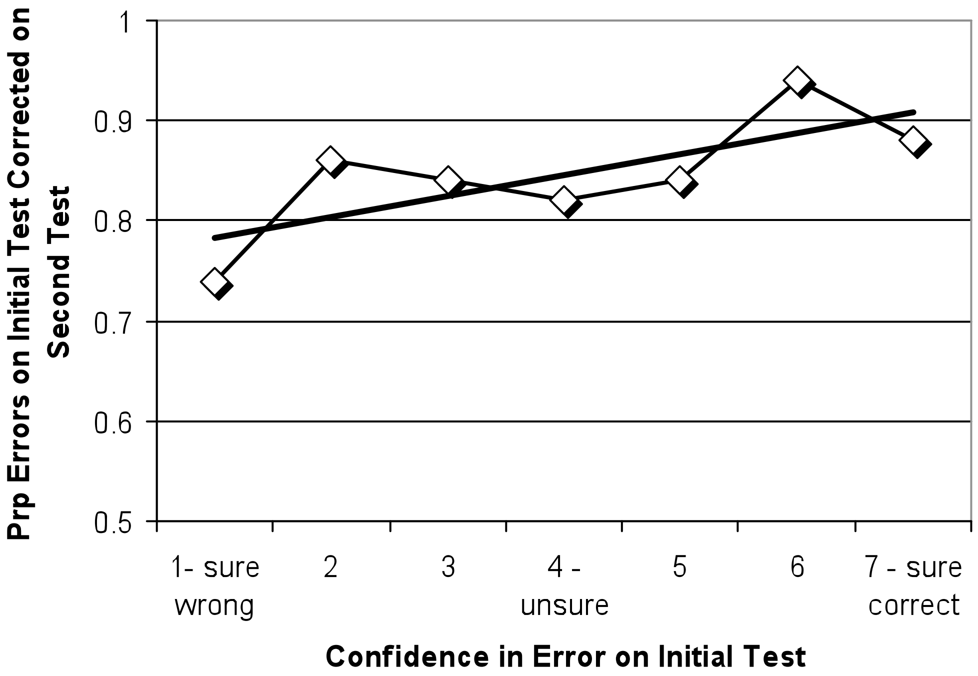


**Figure 10: Average proportion of errors on the initial test that were corrected on the second test, separated by initial confidence rating. Bolded line is the best fitting trend line. (Experiment 5).**

The next question was whether high-confidence errors that were corrected on the second test remained corrected on the third test. That is, did the corrections remain one week later or did the initial high-confidence errors reappear? Thus, this analysis examines only initial errors that were then corrected on the second test. Two subjects were not included in this analysis because all of their initial errors that were corrected on the second test remained correct on the third test. As shown in Figure 11, the high-confidence errors remained corrected on the third test. In fact, initial high-confidence errors were more likely to remain corrected than were low-confidence errors. The mean within-subject gamma correlation was again positive and significant, $\gamma$ = .22, *t*(31) = 2.80, *SEM* = .08.

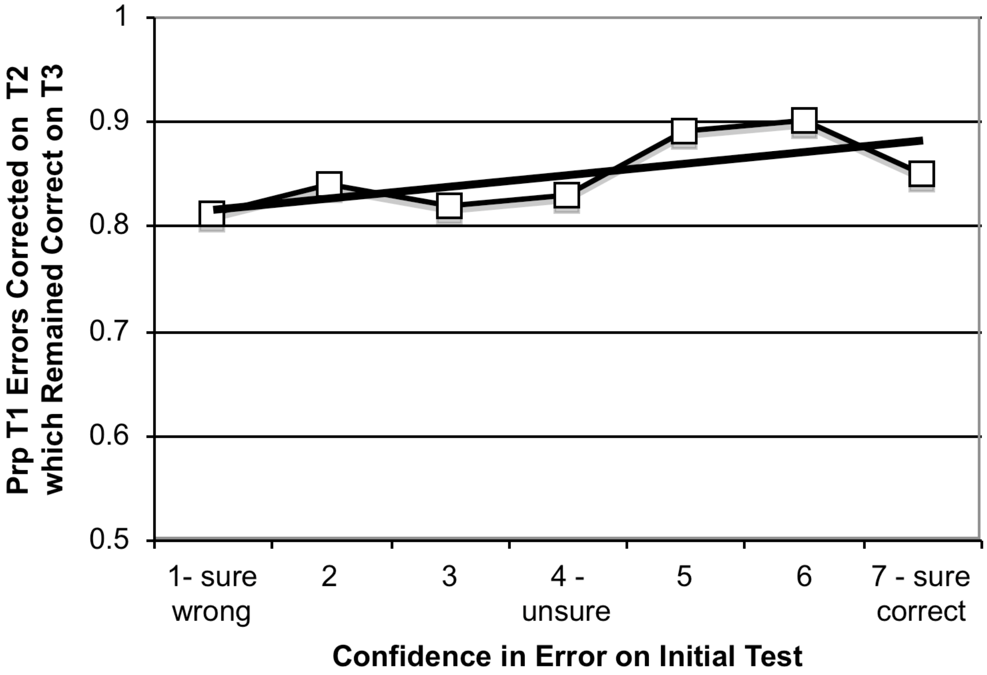


**Figure 11: Average proportion of initial errors corrected on the second test that remained correct on the third test, separated by initial confidence rating. Bolded line is the best fitting trend line. (Experiment 5).**

The last question was whether the high-confidence errors were likely to reappear on the third test after being corrected earlier. That is, after being corrected initially, did subjects' original errors reappear after the delay? Thus, this analysis targets the proportion of initial errors corrected on the second test that then switched back to the subject's original error on the third test. Eight subjects were not included in this analysis because they had no initially corrected errors that switched back to their original error on test three. Overall, very few of the corrected errors reappeared on the third test, and this was unrelated to initial confidence level. As shown in Figure 12, the proportion of initial errors that reappeared on the third test after being corrected on the second test was constant across the different confidence levels. The mean within-subject gamma correlation did not differ from zero, $\gamma = .07$, $t < 1$.

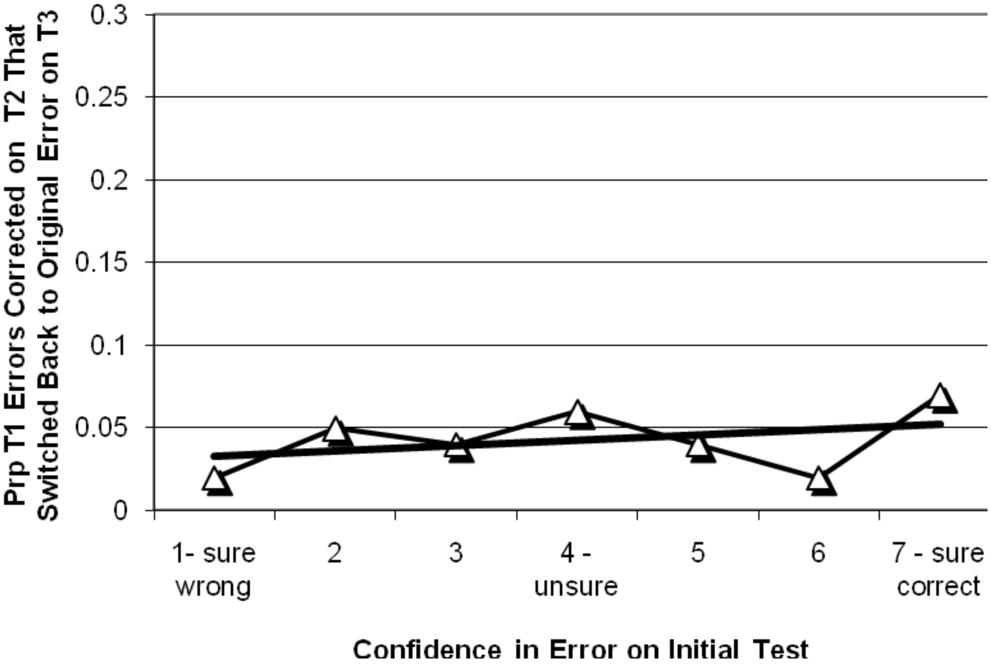


**Figure 12: Average proportion of initial errors corrected on the second test that were incorrect on the third test, separated by initial confidence rating. Bolded line is the best fitting trend line. (Experiment 5).**

## *5.3 Discussion*

These results provide clear evidence that the consequences of the hypercorrection effect persist one week later. Following feedback, subjects were more likely to correct high-confidence errors than low-confidence errors on an immediate test.

These corrected errors then remained corrected one week later. In fact, of the initial errors corrected on the second test, high-confidence errors were more likely to remain corrected than low-confidence errors. Not only are initial high-confidence errors more likely to be corrected on an immediate test, but they are also more likely to stay corrected one week later.

Finally, we found no evidence for spontaneous recovery of the initial high-confidence errors. Only 5% of the time did subjects reproduce their initial error on the third test after correcting the error on the second test. Furthermore, this proportion did not vary with subject's initial confidence in their errors. Recovered incorrect answers on the third test were equally likely for initial low-confidence and high-confidence errors. These findings are consistent with prior results showing that presenting feedback can lead to long-lasting changes in the learner's knowledge (e.g. Pashler et al., 2005). Even after a one-week delay, initial high-confidence errors remained corrected.

One question that remains is why the older and well established high-confidence errors did not reappear after the delay. There are a number of possibilities. Remember that Jost's law states that "if two memories are of equal strength then the older memory will decay at a slower rate than the younger memory" (Wixted, 2004). Thus, one possible explanation is that the initial high-confidence error and the correction are not of equal strength in memory. Given the amount of attentional resources that subjects use to encode surprising feedback (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009), it is possible that the new correct information may be more strongly encoded than the incorrect information. The second possibility is that our delay was not long enough to observe the hypothesized crossover in memory strength. It may be that with a longer delay and greater forgetting more of the initial high-confidence errors would have reappeared on the third test.

A second idea relevant to the current discussion is the "new theory of disuse" proposed by the Bjorks (Bjork, 2001; Bjork & Bjork, 1992). The researchers suggest that memory representations are categorized by two different strengths, storage strength and retrieval strength. Storage strength refers to how well learned and well integrated information is in memory, while retrieval strength refers to how likely a memory is to be retrieved. Storage strength accumulates through both study and retrieval and increases in storage strength are thought to be permanent. Retrieval strength, however, both increases with practice and decreases when other representations are retrieved. Furthermore, it is retrieval strength alone that determines if a memory can be recalled. Key to the theory is the idea that retrieval strength is lost more slowly for memories with greater storage strength. Thus, after a period of disuse, older memories with greater storage strength will be more likely to be recalled than newer memories.

Applying this theory to the current paradigm, the original high-confidence errors should have greater storage strength than the newly learned correct information. Thus, like Jost's law, the new theory of disuse suggests that the original errors should reoccur after an appropriate delay. More experimentation needs to be done with greater delays to determine if this theory's predictions will hold for this paradigm. At the moment, however, we found no evidence to suggest that high-confidence errors are likely to reappear after being initially corrected.

# 6. General Discussion

What have we learned about the hypercorrection effect? First, it is a robust phenomenon and occurs in many different situations. We found that the hypercorrection effect occurs with both episodic and semantic materials. Even high-confidence false memories are more likely to be corrected than low-confidence memory errors. It also occurs both when the feedback is presented immediately after each question and when the feedback is delayed until the end of the test. Furthermore, the effect is not limited to "tricky" questions where one is likely to make a high confidence error. The effect still occurs even when you limit the analysis to questions where most incorrect answers are given with low confidence. Finally, the effect is not transient and persists over a week delay.

Second, we now have further evidence that the effect is due to subjects better encoding unexpected feedback. As discussed earlier, there are two prominent explanations used to understand the hypercorrection effect. The first, the surprise hypothesis suggests that subjects pay more attention to unexpected feedback with benefits for later memory. The second, the domain hypothesis, suggests that high-confidence errors are more likely to be corrected because they occur in domains where the subject already has preexisting knowledge. Thus, the feedback is more easily integrated into memory and is more likely to be remembered on a later test. While both hypotheses are likely true in some situations, the evidence presented here suggests that the domain hypothesis alone cannot explain the hypercorrection effect.

We first examined evidence in favor for the surprise hypothesis. This explanation suggests that feedback should be well attended in two different situations. The first is after a high confidence error as in the typical hypercorrection effect. The second is after a low confidence correct answer. If the hypercorrection effect results from better memory for surprising feedback, then both types of responses should lead to improved memory for the resulting feedback. That is what we found in Experiments 2 and 3. People were more likely to remember details about the feedback such as the color it was presented in and which person presented the feedback after both high confidence errors and low confidence correct answers. In contrast, memory was poor for feedback presented after high confidence correct answers and low confidence errors. When the feedback was expected memory for the feedback was low. We have also presented evidence that the hypercorrection effect occurs in situations where the domain hypothesis is not applicable. In Experiment 4 we found that the surprise hypothesis still occurred with episodic memories. While semantic knowledge is required to understand and remember sentences in episodic memory, there is no reason to believe that the amount of domain knowledge required differs across with subjects 'confidence in their errors. Thus, while domain knowledge may sometimes play a role in the hypercorrection effect, it is not a requirement.

While we have shown the hypercorrection effect to be widespread and robust, it is still unclear how the effect would play out in actual classroom situations. One open question is how effective the hypercorrection effect would be in correcting errors that are more conceptual in nature. In all of the experiments here we have examined individual pieces of information. Correcting an error in these situations involves only changing a single fact and not changing an overall knowledge structure. For example, learning that the capital of Canada is Ottawa, not Toronto, does not require me to reorganize all of my knowledge about Canada. In contrast, children and adults often have errors in their knowledge that are much more deeply ingrained and require more effort to change. For example, when looking at the equation 5 + 3 + 2 = __ many children believe that the equals sign means "add up the numbers to the left" rather than "make the two sides equal."  Thus, the children are unable to solve equations such as 5 + 3 = __ + 2.  Prior research has shown that the children can readily learn the procedure for solving novel problems, but the underlying concept does not change.  When confronted with transfer problems (e.g. 5 +3 = 6 + _) the students again have difficulties (Rittle-Johnson & Alibali, 1999).  Even if the students are highly confident in their incorrect beliefs, it seems unlikely that simply telling them the correct definition will change their larger algebraic concepts.  While feedback alone may not be able to correct these conceptual errors, it may be that actively contrasting student's errors with the correct information will increase correct responding.  The surprise hypothesis would suggest that highlighting the discrepancy between the student's response and the correct answer should help the student to remember the feedback later on.

A separate issue is whether students will use the feedback even if they remember it.  Our experiments have been conducted with students in an artificial situation where the readily believe that the feedback we are giving them is true.  In the schools, however, teachers often have to deal with students who are resistant to what the teacher is saying or who will continue to believe their naïve theories rather than modify them to fit what they are learning in class.  For example, because hats and mittens keep children warm, children often believe that the hat itself creates heat. Even in the face of conflicting evidence students may continue to trust their naïve theories (B. Watson & Kopnicek, 1990).  Feedback is unlikely to be as powerful when the students have reasons to resist its effects.

In conclusion, we have reliably shown that high-confidence errors are more likely to be corrected following feedback than are low-confidence errors.  Furthermore, we have shown that these corrections persist for at least one week and that the effect occurs for both semantic and episodic memories.  In addition, we have shown that the reason these high-confidence errors are likely to be corrected is because students pay more attention to feedback that is discrepant with their expectations.  These findings have obvious implications for education and for correcting students' misconceptions about the world.  As of yet, however, it is unclear whether these results will hold for more complex errors and beliefs.

# References

Abrams, L., Trunk, D. L., & Margolin, S. J. (2007). Resolving tip-of-the-tongue states in young and older adults: The role of phonology. In L. O. Ranal (Ed.), Aging and the Elderly: Psychology, Sociology, and Health. Hauppauge, NY: Nova Science Publishers, Inc.

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. Journal of Educational Research, 62(148-156).

Angell, G. W., & Troyer, M. E. (1948). A new self scoring test device for improving instruction. School and Society, 67, 84-85.

Atkinson, R. C. (1969). Information delay in human learning. Journal of Verbal Learning & Verbal Behavior, 8, 507-511.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. Review of Educational Research, 61, 213-238.

Benjamin, L. T. (1988). A history of teaching machines. American Psychologist, 43, 703-712.

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. Journal of Experimental Psychology: Applied, 5, 438-447.

Birenbaum, M., & Tatsuoka, K. K. (1987). Effects of "on-line" test feedback on the seriousness of subsequent errors. Journal of Educational Measurement, 24(2), 145-155.

Bjork, R. A. (2001). Recency and recovery in human memory. In H. L. Roediger, J. S. Nairne, I. Neath & A. M. Surprenant (Eds.), The nature of remembering: Essays in honor of Robert G. Crowder (pp. 396). Washington, DC: American Psychological Association.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn & R. Shiffrin (Eds.), From learning processes to cognitive processes: Essays in honor of William K. Estes (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Brackbill, Y., & Kappy, M. S. (1962). Delay of reinforcement and retention. Journal of Comparative and Physiological Psychology, 55(14-18).

Brackbill, Y., Wagner, J., & Wilson, D. (1964). Feedback delay and the teaching machine. Psychology in the Schools, 1, 148-156.

Brewer, W. F. (1977). Memory for the pragmatic implications of sentences. Memory & Cognition, 5, 673-678.

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. Journal of Memory & Language, 52, 618-627.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. Journal of Experimental Psychology: Applied, 13(4), 273-281.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: Feedback improves retention of low-confidence correct responses. Journal of Experimental Psychology: Learning, Memory, & Cognition, 34(4), 918-928.

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.

Butterfield, B. (2003). The hypercorrection effect and its neural correlates. Dissertation Abstracts International, 66(05).

Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. Cognitive Brain Research, 17(3), 793-817.

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

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. Metacognition and Learning, 1(1), 69-84.

Dahlgren, D. J. (1998). Impact of knowledge and aging on tip-of-the-tongue rates. Experimental Aging Research, 24, 139-197.

Doerkenson, S., & Shimamura, A. P. (2001). Source memory enhancement for emotional words. Emotion, 1(1), 5-11.

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. Memory, 18, 335-350.

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

Feldhausen, J. F., & Birt, A. (1962). A study of nine methods of presentation of programmed learning material. The Journal of Educational Research, 5, 461-466.

Fritz, C. O., Morris, P. E., Bjork, R. A., Gelman, R., & Wickens, T. D. (2000). When further learning fails: Stability and change following repeated presentation of text. British Journal of Psychology, 91, 493-511.

Gilman, D. A. (1969). Comparison of several feedback methods for correcting errors by computer-assisted instruction. Journal of Educational Psychology, 60(6), 503-508.

Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. Journal of Verbal Learning & Verbal Behavior, 21, 207-219.

Guthrie, J. T. (1971). Feedback and sentence learning. Journal of Verbal Learning & Verbal Behavior, 10, 23-28.

Hanna, G. S. (1976). Effects of total and partial feedback in multiple-choice testing upon learning. The Journal of Educational Research, 69, 202-205.

Jensen, B. T. (1949). An independent-study laboratory using self-scoring tests. Journal of Educational Research, 43, 134-137.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin, 114, 3-28.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. Journal of Memory & Language, 32, 1-24.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254- 284.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.

Kulhavey, R. W. (1977). Feedback in Written Instruction. Review of Educational Research, 47(1), 211-232.

Kulhavey, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. Journal of Educational Psychology, 63, 505-512.

Kulhavey, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. Educational Psychology Review, 1(4), 279- 307.

Kulhavey, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. Contemporary Educational Psychology, 10, 285-291.

Kulhavey, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and Response Confidence. Journal of Educational Psychology, 68(5), 522-528.

Kulhavey, R. W., Yekovick, F. R., & Dyer, J. W. (1976). Feedback and Response Confidence. Journal of Educational Psychology, 68(5), 522-528.

Kulhavy, R. W., Yekovick, F. R., & Dyer, J. W. (1976). Feedback and Response Confidence. Journal of Educational Psychology, 68(5), 522-528.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. Review of Educational Research, 58, 79-97.

Lang, A. J., Craske, M. G., & Bjork, R. A. (1999). Implications of a new theory of disuse for the treatment of emotional disorders. Clinical Psychology: Science and Practice, 6, 80-94.

Lewandowsky, S., Stritzke, W. G. K., Oberauer, K., & Morales, M. (2005). Memory for Fact, Fiction, and Misinformation. Psychological Science, 16(3), 190-195.

Lhyle, K. G., & Kulhavey, R. W. (1987). Feedback processing and error correction. Journal of Educational Psychology, 79, 320-322.

Little, J. K. (1934). Results of use of machines for testing and for drill, upon learning in educational psychology. Journal of Experimental Education, 3, 45-49.

Loftus, E. F. (1979). Reactions to blatantly contradictory information. Memory & Cognition, 7, 368-374.

MacKay, D. G., & Ahmetzanov, M. V. (2005). Emotion, Memory, and Attention in the Taboo Stroop Paradigm An Experimental Analogue of Flashbulb Memories. Psychological Science, 16(1), 25-32.

Marsh, E. J., Lozito, J., Bjork, E. L., & Bjork, R. (submitted). Correcting errors and misconceptions: How should feedback be provided during multiple-choice testing?

McConnell, M. D., & Hunt, R. R. (2007). Can false memories be corrected by feedback in the DRM paradigm? Memory & Cognition, 35(5), 999-1006.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. Contemporary Educational Psychology, 16, 192-201.

McDermott, K. B. (1996). The persistence of false memories in list recall. Journal of Memory & Language, 35, 212-230.

McDermott, K. B., & Chan, J. C. K. (2006). Effects of repetition on memory for pragmatic inferences. Memory & Cognition, 34(6), 1273-1284.

McDermott, K. B., & Roediger, H. L., III. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. Journal of Memory & Language, 39, 508-520.

McGeogh, J. A. (1942). The psychology of human learning. New York: Longmans, Green.

Moore, J. W., & Smith, W. I. (1961). Knowledge of results in self-teaching spelling. Psychological Reports, 9, 717-726.

More, A. J. (1969). Delay of feedback and the acquisition and retention of verbal materials in the classroom. Journal of Educational Psychology, 60, 339-342.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. Instructional Science, 32, 99-113.

Morgan, C. L., & Morgan, L. (1935). Effects of immediate awareness of success and failure upon objective examination scores. Journal of Experimental Education, 4, 63-66.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), Handbook of research on educational communications and technology (pp. 745-783). Mahwah, NJ: Erlbaum.

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 26, pp. 125-173). New York: Academic Press.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Roher, D. (2005). When does feedback facilitate learning of words. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1), 3-8.

Pavlov, I. P. (1927). Conditioned reflexes. Oxford: Oxford University Press.

Peeck, J., & Tillema, H. H. (1978). Delay of feedback and retention of correct and incorrect responses. The Journal of Experimental Education, 47(2), 171-178.

Peeck, J., van den Bosch, A. J., & Kreupeling, W. J. (1985). Effects of informative feedback in relation to retention of initial responses. Contemporary Educational Psychology, 10, 303-313.

Phye, G. D., & Andre, T. (1989). Delayed Retention Effect: Attention, perseveration, or both? Contemporary Educational Psychology, 14, 173-185.

Pressey, S. L. (1926). A simple apparatus which gives tests and scores -- and teaches. School and Society, 23, 373-376.

Pressey, S. L. (1927). A machine for automatic teaching of drill material. School and Society, 25, 549-552.

Preston, R., & Preston, S. (2005). The Amazing 10,000 Quiz Challenge. Buffalo: Firefly.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. Psychological Review, 88, 93-134.

Renner, K. E. (1964). Delay of reinforcement: A historical review. Psychological Bulletin, 61, 341-361.

Rescorla, R. A. (2004). Spontaneous recovery. Learning & Memory, 11, 501-509.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical Conditioning II: Current Research and Theory (pp. 64-99). New York: Appleton-Century-Crofts.

Ripple, R. E. (1963). Comparison of the effectiveness of a programmed text with three other methods of presentation. Psychological Reports, 12, 227-237.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? Journal of Educational Psychology, 91, 175-189.

Roper, W. J. (1977). Feedback in computer assisted instruction. Programmed Learning and Educational Technology, 14, 43-49.

Skinner, B. F. (1954). The science of learning and the art of teaching. Harvard Educational Review, 24, 86-97.

Skinner, B. F. (1958). Teaching Machines. Science, 128, 969-797.

Sturges, P. T. (1972). Information delay and retention: Effect of information in feedback and tests. Journal of Educational Psychology, 63, 32-43.

Thorndike, E. L. (1911). Animall Intellegence. Experimental Studies. Oxford, England: Macmillian.

Thorndike, E. L. (1913). Educational Psychology. New York: Teacher's College, Columbia University.

Tsai, C. I., Klayman, J., & Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. Organizational Behavior and Human Decision Processes, 107, 97-105.

Vigliocco, G., Vinson, D., Martin, R., & Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of tip of the tongue states and anomia. Journal of Memory & Language, 40, 534-558.

Watson, B., & Kopnicek, R. (1990). Teaching for conceptual change: Confronting children's experience. Phi Delta Kappan, 680-684.

Watson, J. M., McDermott, K. B., & Balota, D. A. (2004). Attempting to avoid false memories in the Deese/Roediger-McDermott paradigm: Assessing the combined influence of practice and warnings in young and old adults. Memory & Cognition, 32(1), 135-141.

White, K. (1968). Delay of test information feedback and learning in a conventional classroom. Psychology in the Schools, 5, 78-81.

Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. Psychological Review, 111(4), 864-879.

Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible information: Potential bases for the incubation effect in problem solving. Journal of Experimental Psychology: Learning, Memory, & Cognition, 13, 187-205.

# Biography

**Born**

April, 4th 1982 in Bloomington, IN

**Education**

Duke University, Durham, NC
Ph.D., Psychology, 2010

Duke University, Durham, NC
M.A., Psychology, 2006

Washington University in St. Louis, St. Louis, MO
B. A., Psychology, 2004

**Publications**

Fazio, L. K., & Marsh, E. J. (in press). Correcting false memories. *Psychological Science*.

Fazio, L. K., Agarwal, P. K., Marsh, E. J. & Roediger, H. L., III (in press). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition.*

Fazio, L. K., Huelser, B. J., Johnson, A. & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory,18,* 335-350.

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

Fazio, L. K., & Marsh, E. J (2008). Slowing presentation speed increases illusions of knowledge. *Psychonomic Bulletin and Review, 15,* 180-185.

Fazio, L. K., & Marsh, E. J. (2008). Older, not younger, children learn more false facts from stories. *Cognition, 106,* 1081-1089.

Marsh, E. J., & Fazio, L. K. (2006). Learning errors from fiction: Difficulties in reducing reliance on fictional stories. *Memory & Cognition*, *34,* 1140-1149.

**Honors and Awards**

| | |
|---|---|
| 2008-2009 | Carolina Consortium on Human Development Predoctoral Fellow |
| 2008 | APA Dissertation Research Award |
| 2004 – 2008 | James B. Duke Graduate Fellowship Recipient |
| 2004 | NSF Graduate Fellowship – Honorable Mention |