Journal of Applied Research in Memory and Cognition 3 (2014) 189–197

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors

Janet Metcalfe[a,*], David B. Miele[b]

[a] Columbia University, United States
[b] Boston College, United States

ARTICLE INFO

Article history:
Received 24 November 2013
Received in revised form 28 March 2014
Accepted 1 April 2014
Available online 13 April 2014

Keywords:
Hypercorrection
Error correction
Testing benefits
Metacognition
Confidence judgments
Updating

ABSTRACT

How people correct their mistakes and sustain those corrections over time is a problem of central interest to education. It might be thought that the erroneous beliefs that people hold with high confidence would be especially difficult to correct. Interestingly, people correct these high confidence errors more easily than low confidence errors, a phenomenon known as the 'hypercorrection effect'. Unfortunately, though, with a delay in testing there is a tendency for some of these high confidence errors to reemerge – a finding with serious consequences for education. This study investigated the effect of intervening a test immediately after corrective feedback on preventing the return of the errors. It also investigated processing differences between prevention-focused and promotion-focused people. The most educationally important new finding was that testing immediately after corrective feedback not only greatly enhanced memory for the correct answers but also blocked the return of the errors.

Although it is often asserted that the propensity to commit errors is a fundamental characteristic of being human, correcting those same errors can be seen as one of the primary tasks of education. Whether, as an educator, one should encourage students to express their errors has been a topic of recent debate. A number of research findings indicate that generating errors – as long as corrective feedback is provided (see Butler & Roediger, 2008; Finn & Metcalfe, 2010; Kang, McDermott, & Roediger, 2007; Lhyle & Kulhavy, 1987; McDaniel & Fisher, 1991; Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005) – may help rather than hinder learning (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Kornell, Hays, & Bjork, 2009; Richland, Kornell & Kao, 2009; Slamecka & Fevreiski, 1983). However, in the studies that have demonstrated that making errors helps later memory for the correct answer, the 'errors' that have been produced have nearly always been mere guesses rather than genuine errors that the person believes to be correct. In contrast, the present paper investigates a paradigm in which the errors that people make are genuine, but fallacious, responses to questions of fact. Our focus will be on error correction when people strongly believe that the error that they produced was the correct answer, in contrast to mere guesses or responses about which they expressed low confidence.

It may seem intuitive, and a number of theories of memory support the idea (Barnes & Underwood, 1959; Murdock, 1974; Raaijmakers & Shiffrin, 1981), that the errors that an individual strongly believes are correct should be resistant to correction. For example, classic interference theory proposed that when a cue or question A, is associated with a response B, and then the individual is asked to learn a new response, C, to that original cue, the two responses, B and C compete with one another. The stronger the first response (B), the more difficult it should be to supplant it with a new response (C). Although this logic has usually been applied to experimentally learned responses, and not to the educational issue of correcting erroneous prior knowledge, nevertheless the application to error correction seems appropriate. Despite the plausibility of the conjecture that strong prior (but erroneous) responses should make updating difficult, experimental investigations in which a person is asked a factual question, gives an answer along with his or her confidence in the correctness of the answer, is provided with corrective feedback, and then is retested for the correct answer, have shown that high confidence errors are particularly *easy* to correct, a phenomenon known as the *hypercorrection effect* (Butterfield & Metcalfe, 2001).

This effect has been demonstrated many times with immediate testing (Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008; Eich, Stern & Metcalfe, 2013; Fazio & Marsh, 2009a, 2010; Kulhavy &

* Corresponding author at: Department of Psychology, Columbia University, New York, NY 10027, United States. Tel.: +1 212 854 7971.
E-mail address: jm348@columbia.edu (J. Metcalfe).

http://dx.doi.org/10.1016/j.jarmac.2014.04.001
2211-3681/

Stock, 1989; Metcalfe, Butterfield, Habeck, & Stern, 2012; Metcalfe & Finn, 2012a; Sitzman, Rhodes, & Tauber, 2014). Both children and college-aged adults hypercorrect (Metcalfe & Finn, 2012b), although there is some evidence that older adults may hypercorrect less than college-aged participants do (Eich et al., 2013). Hypercorrection occurs with different types of materials (e.g., Fazio & Marsh, 2010, though see Sitzman & Rhodes, 2010) but has most frequently been studied with general information questions. It also appears to occur with delayed testing, although only two studies involving a delay have, so far, been published (Butterfield & Mangels, 2003; Butler, Fazio, & Marsh, 2011). One of these studies also showed that there is a *tendency for the errors to return at a delay* (Butler et al., 2011) – a phenomenon of central concern educationally and in the present article. We will investigate a potential method to prevent this return of the errors.

Two explanations of the hypercorrection effect have garnered empirical support. The first explanation is that when individuals receive feedback that indicates that the answer they have just given with considerable confidence is wrong, they rally their attentional resources – perhaps because they were surprised, embarrassed or upset at having made the mistake – and devote those resources to efforts to impress into memory the correct answer. When they are wrong but with less confidence, they are less concerned and their attentional response to the feedback is less intense.

Several studies provide support for this explanation. Butterfield and Metcalfe (2006) used a simultaneous tone detection task while people were answering and receiving feedback to high and low confidence errors. Failure to detect a soft tone while the feedback was being presented was taken as indicating that the individual's attention was focused on processing the feedback rather than listening for the tone. People missed the tone presented with the feedback more frequently when they had made a high rather than a low confidence error. An experiment by Butterfield and Mangels (2003) indicates a similar conclusion. They conducted an event related potential study of the hypercorrection effect. Time locking to the onset of the corrective feedback, they found an exaggerated p3 voltage deflection during feedback to high as compared to low confidence errors – again suggesting increased attention to the high confident error corrections. And, finally, Fazio and Marsh (2010) showed that people remembered not only the answer but the background on which the answer was presented, again suggesting that they paid more attention to the entire subsequent event, when they had made a high confidence rather than a low confidence error.

The second, non-mutually exclusive explanation that has been extensively explored with regard to the hypercorrection effect is semantic in nature. People are more familiar with the questions related to errors that are committed with high than with low confidence, as well as to the answers associated with those questions (Butterfield & Mangels, 2003; Butterfield & Metcalfe, 2006). Metcalfe and Finn (2012a) conducted a latent semantic analysis (LSA, Landauer & Dumais, 1997) of the relatedness of the errors to the correct answers, both for high and for low confidence errors. This analysis revealed a tighter associative relation between the target and high confidence errors than between the target and low confidence errors. Furthermore, both children and adults claim, upon being given the correct responses, that they 'knew the answers all along' (Metcalfe & Finn, 2012a,b) more for high than for low confidence errors. This is not purely a hindsight bias (Fischhoff, 1975; Hawkins & Hastie, 1990). Metcalfe and Finn (2012a) showed that if before being given the corrective feedback people are asked to make a second guess, or to choose the correct answer in a multiple choice test that does not include their mistake, or they are given successive clues about the answer, they are more likely to produce the answer, to correctly select the answer, and they need fewer clues, for high confidence errors as compared to low confidence errors. These data indicate that they did, in fact, know (all along) something more about the answers to high than to low confidence errors.

It seems likely that while both of these factors are interactive (see Sitzman et al., 2014) and contribute to hypercorrection, the attentional factor might be more short lived than the semantic structure factor. Furthermore, people's semantic structure might be subject – over long delays – to a regression toward its pre-corrective feedback state. Such a regression to the pre-experimental state of semantic memory – a state which included the erroneous response as correct – could presage the return of the high confidence errors over time. This phenomenon was observed by Butler et al. (2011) and is the prime concern of the present article. While these researchers showed that the hypercorrection of high confidence errors persisted at a delay of over a week, they also demonstrated that there was a tendency for some of the original high confidence errors to re-emerge. They suggested that these errors might *not* reemerge if an intervening test were imposed, citing an unpublished study by Fazio and Marsh (2009b) that failed to show evidence for the return of the errors at a delay. However, the failure to find a return of errors, in isolation, is a null effect that could have resulted from myriad causes. While potentially of great educational interest, then, whether testing immediately following corrective feedback might be sufficient to prevent the return of the errors is currently unknown. In the present experiment, we directly test this conjecture in a well controlled within-participants design.

There is a third potential explanation – the recursive reminding hypothesis – that has not previously been applied to the hypercorrection effect but that also deserves attention. Wahlheim and Jacoby (2013, and see Jacoby & Wahlheim, 2013) have proposed that, rather than proactive inhibition (as would be expected by interference theory, as discussed above), proactive *facilitation* will occur if (a) the participant notices at the time of presentation of C that there is a change from the earlier B response, and if (b) they later 'bring to mind' (Jacoby, Wahlheim, & Yonelinas, 2013, p. 638) the B item at time of test for C (and see, Hintzman, 2004). If, however, participants do not bring item B to mind at time of test, its earlier occurrence in the context of A results in proactive interference. In the hypercorrection context, B is the original error, which, in the experimental procedure that follows is corrected the moment it is committed. It is very unlikely that participants fail to notice the change from the error to the correction at time of presentation of the correct answer, given that there is no time delay between the generation of the erroneous response and the presentation of the correction. It is not currently known, however, whether people bring to mind their original high confidence errors more than their low confidence errors when they are tested. But it seems likely. It is also unknown, in the context of factual error correction, whether bringing to mind the original errors helps, rather than hurts (or leaves unaffected) memory for the corrections. These are empirical questions. In the experiment that follows, to investigate whether the original errors come to mind, and whether doing so helps or hurts, we asked people to produce two answers to each question at time of post feedback recall and to indicate which of the two is correct. The recursive reminding hypothesis, to explain the hypercorrection effect, would predict that (a) the original errors should show up more for high than for low confidence errors, and that (b) the probability of error correction, that is of producing the correct 'C' responses and knowing that they are correct, should be higher when the original errors come to mind than when they do not.

This paper also investigates the possibility that there might be individual differences in error correction dependent upon what Higgins (1997), and colleagues, refer to as "regulatory focus." According to regulatory focus theory (Higgins, 1997; see Molden & Miele, 2008, for a review), people who are primarily *promotion-focused*, tend to eagerly seek opportunities for gain that will move them closer to their goals. As a consequence, they are willing to take chances and are relatively unconcerned that in doing so they may make errors. If they do make errors, they tend to brush them off, and get on with their task of seeking opportunities. A signal detection analysis of such participants' performance on a yes-no recognition memory task indicated that they exhibited a low decision criterion (Crowe & Higgins, 1997; cf. Scholer, Stroessner, & Higgins, 2008). In contrast, people who are *prevention-focused* tend to vigilantly protect against potential losses and to be concerned about mistakes. They exhibited a high criterion on the recognition memory task. Because prevention-focused and promotion-focused individuals might deal with errors differently we decided to measure this variable using the regulatory focus questionnaire (RFQ) devised by Higgins et al. (2001). Although a number of studies have shown that the effects of positive and negative (non-corrective) feedback are moderated by regulatory focus (e.g., Idson & Higgins, 2000; Shu & Lam, 2011; Van-Dijk & Kluger, 2004), we know of no previous studies that have examined promotion and prevention focus differences in relation to semantic memory errors followed by presentation of the correct answers (i.e., *corrective* feedback).

Prevention and promotion-focused people seem likely to differ on the attentional factor that contributes to the hypercorrection effect. If prevention focused people's investment in, and caring about mistakes translates into increases in attention, then we would expect them to ramp up their attention to prevent further errors. Since, in general, only the high confidence mistakes, and not the guesses, are considered to even be genuine errors, the differential effect should focus on the high confidence errors. Accordingly, it seemed plausible that prevention-focused participants might show a greater attentionally based hypercorrection effect than promotion-focused participants. Although the second, semantic, factor also contributes to the hypercorrection effect, it seemed unlikely to us that there would be a difference between promotion- and prevention-focused college students; nevertheless, we accounted for this factor using a latent semantic analysis.

Although this study sought primarily to determine whether we could find a way to keep the high confidence errors that people committed from coming back, we secondarily investigated (a) the impact of the mental presence of prior errors on error correction, (b) whether regulatory focus had an effect on error correction, and (c) the reliability of the hypercorrection effect itself at a delay. Assuming that we would replicate the finding of Butler et al., 2011 – that there was a tendency for the high confidence errors to return over time – our primary question was whether we could prevent this from happening by intervening a test immediately following corrective feedback? Our hypothesis was that testing would not only bolster correct recall, as has often been found (Roediger & Karpicke, 2006a,b), but that it would also ward off the return of errors. To test this hypothesis, the experiment that follows includes an immediate testing condition, as well as two delayed testing conditions: one that is preceded by the immediate test, and one that is not.

1. Method

1.1. Participants

Participants were 50 Columbia University or Barnard College students who were either paid or who received course credit for their participation.[1] A sensitivity analysis conducted using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) showed that this sample size (49 after exclusions; see below) provides approximately 80% power for detecting a medium to large correlation (.39) between our measure of regulatory focus and a given dependent variable. The sample included 21 males and 29 females; mean age was 20.72 (*SD* = 3.48). Participants were treated in accordance with the standards of the American Psychological Association, and all procedures were reviewed and approved by the Columbia University IRB.

1.2. Design and Procedure

In the pretest, participants were presented, one at a time on the computer, with a series of individually randomized general information questions taken from a list of 439 such questions of varying levels of difficulty. The list represented a subset of questions from a larger pool (see Butterfield & Metcalfe, 2006) that included questions from Nelson and Narens (1980), various board games, and internet trivia sites. A sample question is "What igneous rock makes up the bulk of Devils Tower?" (answer: "basalt"). All correct answers consisted of a single word. After the presentation of each question, the participant entered his or her response, followed by a confidence judgment about the correctness of the response just given (both of these actions were self-paced and participants were allowed to skip to the confidence judgment without entering a response). The judgment was made using a horizontal analog slider that ranged from "not confident" to "very confident"; each judgment was automatically coded on a 0–100 scale by the computer. Then, regardless of the response given, the correct answer was provided by the computer and remained onscreen until the participant pressed a button to continue to the next question. If the participant's response was determined to be correct by the computer, the answer was displayed in a green font; if the response was incorrect, the answer was shown in a red font.[2] The computer continued presenting questions, recording responses and confidence judgments, and providing feedback until the participant had answered 200 questions incorrectly, had gone through all but one of the questions, or 30 min had passed – whichever came first.

The computer then informed participants that they would take a recall test. For the test, the participants' incorrect responses were partitioned into five groups based on the confidence ratings they gave for those responses (0–24, 25–49, 50, 51–75, and 76–100). Approximately half of the responses in each group were selected for the immediate test and were presented in a random order. For each of these questions, participants were given as much time as they needed to enter the first two responses that came to mind and to then click a button next to the response that they thought was most likely to be correct. They were not allowed to proceed unless they had entered text into both response boxes and made a selection. Once they submitted their responses, they were given as much time as they needed to make a remember/know/guess judgment for their selected response (the results of this judgment are not reported here). Having participants list two responses for each question allowed us to investigate issues related to the mental presence of the original error during testing. After participants responded to each question, the computer did *not* provide feedback, as this set of questions would be presented again during the second session.

The participants returned to the lab, at least one week later (most participants returned exactly 7 days later, but a number of them returned from 8 to 14 days later) for a test on the questions they had answered incorrectly during the pretest. They were again asked to provide two responses to each question and to indicate which of the responses was correct. During the delayed test, *all* of the questions that participants had answered incorrectly

[1] This total does not include 13 participants whose Session 1 data were lost due to a computer error.

[2] The data were later examined by a research assistant in order to identify trials in which the computer mistakenly indicated that the participant's response was incorrect. These trials were excluded from analyses.

during the pretest were presented in a random order. Note that half of these questions had already been presented during the immediate test, whereas the other half had not been tested since the pretest. These will be called the *retested* and *untested* delay conditions, respectively. The delayed test followed the same procedure as the immediate test, with the exception that participants received feedback about the correct answer (in a black font) after making the remember/know/guess judgment.

After completing the general information task, participants filled out the Regulatory Focus Questionnaire (RFQ; Higgins et al., 2001; http://www.columbia.edu/cu/psychology/higgins/papers/rfq.pdf), which includes 11 items and assesses their subjective history of promotion and prevention success. The RFQ has been shown to be one of the most widely used, valid and reliable measures of chronic (i.e., dispositional) regulatory focus (Haws, Dholakia, & Bearden, 2010). They compared the RFQ to several other measures of chronic regulatory focus and concluded that "Of the five measures, it alone is adequate in internal consistency, homogeneity, and stability, and it performs the best in terms of predictive validity and representativeness. Notably, it is the only scale to distinguish explicitly between approach and avoidance in each regulatory focus and to encapsulate the key tenets of regulatory focus theory" (p. 979). After filling out the RFQ, participants completed a suspicion check and a demographic form and were then thanked and debriefed.

2. Results

One participant spent an abnormally long time viewing feedback on questions answered incorrectly on the pretest (over 6 *SD* above the mean). Data from this participant were excluded from analyses.

To assess individual differences in participants' regulatory focus, we averaged the six promotion items (after reverse coding where necessary; $M = 3.61$, $SD = .52$, range = 2.00–4.83, $\alpha = .50$) and the five prevention items ($M = 3.30$, $SD = .71$, range = 1.20–5.00, $\alpha = .74$). Because promotion and prevention motivations are thought to vary independently (such that an individual can be high on both, one, or neither), we computed each participant's predominant regulatory focus by subtracting their prevention score from their promotion score, consistent with previous regulatory focus studies (e.g., Bohns et al., 2013; Cesario, Grant, & Higgins, 2004; Hong & Lee, 2008; Scholer, Ozaki, & Higgins, 2014). We then transformed this variable ($M = .31$, $SD = .91$, range = −1.90 to 1.97, skewness = −.61, kurtosis = .11) to be centered at the midpoint of the index (such that 0 indicated no predominance), with a standard deviation of 1. This index of participants' regulatory focus was treated as a *continuous variable* in all analyses. A positive correlation between the index and a dependent variable indicated that the more promotion-focused and less prevention-focused participants were, the higher they scored on the dependent variable. When estimating effects for promotion-focused participants, we conducted analyses at 1.5 *SD* above the midpoint of the index, and when estimating effects for prevention-focused participants, we conducted analyses at 1.5 *SD* below the midpoint of the index.

2.1. Basic results

Participants completed over 120 questions on average during the pretest (it appears that all participants stopped after reaching the 30 min time limit), of which 32.61 ($SD = 14.97$) were correct, 62.14 ($SD = 23.96$) were errors of commission, and 26.53 ($SD = 35.10$) were errors of omission. There was no association between the regulatory focus index and how many questions participants completed overall, how many they answered correctly,

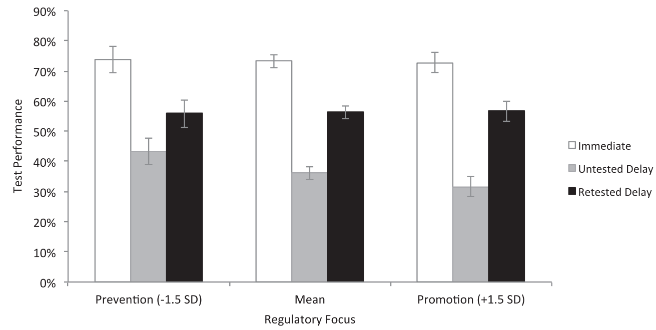

Fig. 1. Percent correct as a function of testing condition and regulatory focus. Means for promotion and prevention are estimated at 1.5 *SD* above and below the *midpoint* of the regulatory focus index, respectively. Error bars reflect standard errors of the mean.

how many they answered incorrectly (commission errors), or how many they declined to answer (omission errors, *r*s ranged from −.08 to .07, $ps > .60$). For all of the analyses, we excluded the trials in which participants omitted responses on the pretest.

The mean gamma correlation (.78, $SD = .10$) between confidence and correctness on the pretest was strongly positive. This indicates that most of the time the participants knew when they were right and when they were wrong. There was no association between this gamma correlation and the regulatory focus index ($r = .08$, $p = .59$).

The distribution of confidence on the errors that were produced was divided into four bins: 0–25, 26–50, 51–75 and 76–100 on the confidence scale (although not all participants used the full scale when making confidence judgments), with mean counts of 41.78 ($SD = 23.22$), 8.33 ($SD = 6.35$), 7.57 ($SD = 6.92$), 4.47 ($SD = 4.50$), respectively. The distribution did not vary by regulatory focus (*r*s ranged from .01 to .06 for each quartile, $ps > .71$).

2.2. Correct recall in the three test conditions

When analyzing data from the three tests that followed corrective feedback on the errors, we followed a two step procedure. In the first step, we conducted a repeated measures ANOVA with test condition (immediate vs. retested delay vs. untested delay) as the independent variable. In the second step, we added the regulatory focus index as a covariate and conducted a repeated measures ANCOVA. We report the main effect of test condition from the first step, and the main effect of regulatory focus, as well as the regulatory focus × test condition interaction, from the second step. To explore the pattern of interactions, we conducted an uncorrected set of correlations between regulatory focus and the dependent variable within each test condition.

As shown in Fig. 1, there was a main effect of test condition on correct recall, $F(2, 96) = 229.37$, $p < .001$, $\eta^2_p = .83$. Participants performed significantly better on the immediate test than on either of the delayed tests ($ps < .001$), and significantly better on the delayed test for retested items than on the delayed test for previously untested items ($p < .001$). These results were expected. In the immediate test, corrective feedback had been provided only a few minutes earlier, resulting in a high level of performance. The fact that performance in the retested delayed condition was better than in the untested delayed condition is a classic test effect (Roediger & Karpicke, 2006a,b). But even the answers to items that did not benefit from the immediate test, but were given only a single instance of corrective feedback, were well remembered over the week, rising from 0% to about 36% correct performance with one very brief exposure of the correct answer.

There was a significant regulatory focus × test condition interaction, $F(2, 94) = 3.51$, $p = .03$, $\eta^2_p = .07$ (Greenhouse–Geisser correction: $p = .04$). Uncorrected post hoc tests directed at this

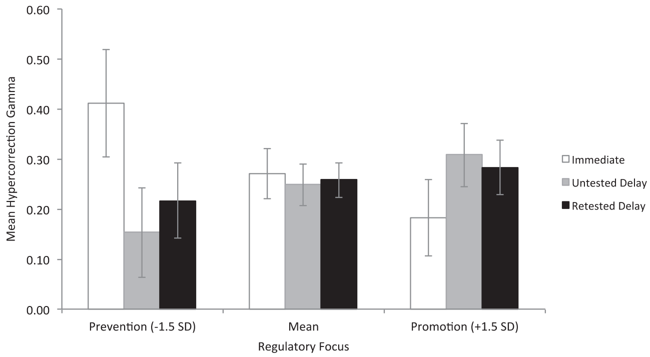

Fig. 2. The hypercorrection effect immediately and at the two kinds of delay, for prevention and promotion-focused participants. Means for promotion and prevention are estimated at 1.5 *SD* above and below the *midpoint* of the regulatory focus index, respectively. Error bars reflect standard errors of the mean.

significant interaction revealed that prevention-focused participants performed marginally better than did promotion-focused participants in the delayed test for the previously untested items ($r = -.26$, $p = .08$), but not differently in the immediate test or retested delayed test condition (*r*s ranged from −.03 to .02, *ps* > .87).

2.3. Hypercorrection

Significant hypercorrection was found in all conditions of this experiment, as is illustrated in Fig. 2. The gamma correlation between original confidence in errors and the correctness of response was significantly greater than zero for each of the three criterion tests (immediate, untested delay, retested delay; *ts* > 5.30, *ps* < .001), and did not vary by condition, $F < 1$.[3]

Although there was no main effect of regulatory focus on the gamma correlations, $F < 1$, the hypercorrection interaction between regulatory focus and test delay was significant, $F(2, 92) = 3.07$, $p = .05$, $\eta^2_p = .06$, as is apparent from Fig. 2. However, uncorrected post hoc tests indicated that the correlations between the regulatory index and hypercorrection gammas in each of the three testing conditions were not significant (*r*s ranged from −.22 to .18, *ps* > .14). Uncorrected pair-wise comparisons were conducted – one set for the means in each test condition estimated at 1.5 *SD* above the midpoint of the regulatory focus index (i.e., for promotion-focused participants) and another set for the means estimated at 1.5 *SD* below the midpoint of the regulatory focus index (i.e., for prevention-focused participants). While none of the between-test comparisons for promotion-focused participants was significant (*ps* > .18), the hypercorrection gammas for prevention focused participants were higher in the immediate condition than were the gammas in the two delayed conditions (the comparison was at the threshold of significance for the untested delay condition, $p = .05$, and marginally significant for the retested delay condition, $p = .07$). These results suggest that the prevention focus participants did selectively attend to the high confidence errors, resulting in an exacerbated hypercorrection effect on the immediate test, but that this attentional effect did not persist at a delay.

2.4. Latent semantic analysis

This analysis allowed us to investigate (a) whether there were semantic memory differences between high and low confidence errors and their corrections, and (b) whether or not there were differences between prevention and promotion focused individuals with respect to the structure of their semantic memories for errors. LSA is method that extracts contextual-usage ‘meaning’ of words via statistical computations applied to a large corpus of text. It can be used to evaluate the similarity of the meanings of words, as given by the cosines (Landauer & Dumais, 1997; http://cwl-projects.cogsci.rpi.edu/msr/). We computed the cosines between the high confidence errors and the correct responses on the pretest and between the low confidence errors and the correct responses to investigate whether there were differences between the promotion- and prevention-focused participants that might have accounted for the hypercorrection differences observed across the test conditions. There was a significant effect of confidence (errors that had a confidence judgment of over 50 on the slider scale versus errors that had a judgment of 50 or under) on the cosines, $F(1, 48) = 14.49$, $p < .001$, $\eta^2_p = .23$, such that the high confidence errors ($M = .25$, $SD = .10$) had a stronger semantic association with the answer than the low confidence errors ($M = .20$, $SD = .06$). Neither the main effect of regulatory focus, nor the regulatory focus × test condition interaction, both *F*'s < 1, was significant.

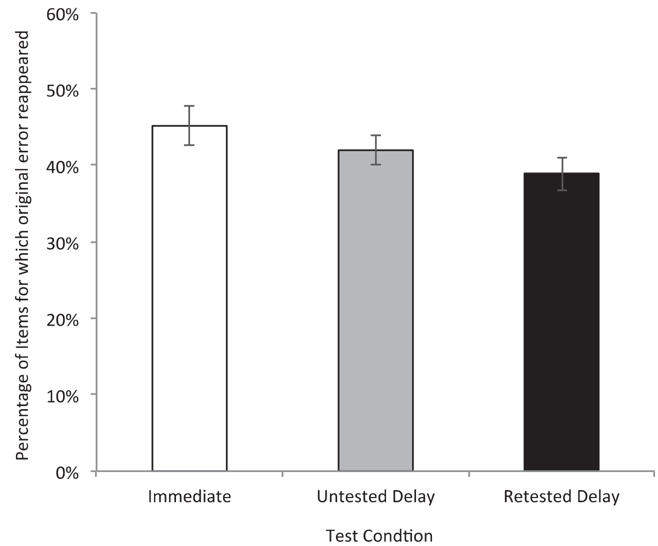

Fig. 3. Show up of the original errors as one of the two responses given. Error bars reflect standard errors of the mean.

2.5. Return of the errors

2.5.1. Show up of errors regardless of ascription of correctness

To address the question of whether the errors tended to show up or become mentally present at a delay, we first computed the probability of the original error being either one of the two responses that the person was requested to give for each question, regardless of whether they ascribed the response as being correct. This analysis indicates the rate at which the errors showed up, regardless of whether they were or were not thought to be correct. As shown in Fig. 3, there was a significant main effect, $F(2, 96) = 4.03$, $p = .02$, $\eta^2_p = .08$ (Greenhouse–Geisser correction: $p = .03$), such that the presence of the error was significantly greater in the immediate condition than in the retested delay condition ($p = .005$). The difference between the immediate condition and the untested delay condition was not significant ($p = .22$), nor was the difference between the retested and the untested delay condition significant ($p = .11$). Neither the main effect of regulatory focus on show up of the errors, $F(1, 47) = 1.57$, $p = .22$, $\eta^2_p = .03$, nor the regulatory focus × test condition interaction, $F(2, 94) = 2.06$, $p = .13$, $\eta^2_p = .04$, was significant.

As is relevant to the recursive reminding hypothesis, we also computed whether the show up of the original errors was greater

[3] A gamma correlation could not be computed for one participant in the immediate test condition.

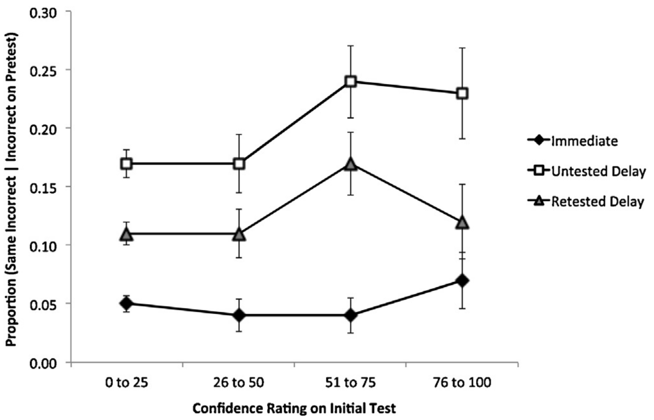

Fig. 4. Errors that returned on a later test and were ascribed as being correct, as a function of original confidence in the error, collapsed over all participants. Error bars reflect standard errors of the mean.

for high than for low confidence errors, by computing gamma correlations in each condition between error show up and original confidence in the error. Those gammas were .28 for the immediate condition, .32 for the untested delay condition, and .27 for the retested delay condition. The gammas were all greater than zero ($ts > 6.77$, $ps < .001$) and did not very by condition ($F < 1$). There was also no main effect of regulatory focus in the gammas, nor was there a regulatory focus x condition interaction ($Fs < 1$). The fact that in all three conditions the show up of the original errors was greater for high than for low confidence errors satisfies the first condition of the potential applicability of the recursive reminding hypothesis discussed in the Introduction.

2.5.2. Return of the errors: ascription of correctness to original high confidence errors

We investigated people's (mistaken) ascriptions that the original high as compared to low confidence errors that showed up were correct. Butler et al. (2011) had reported this as 'errors returning.' We used the same analyses that Butler et al. (2011) had used. The results, shown in Fig. 4, represent the means collapsed across all of the trials, divided into confidence quadrants. They do not take into account participant means, because many of the participants did not have data points at each of the confidence intervals. The pattern in this figure suggests that high confidence errors might have returned, selectively, in the untested delayed condition and not in the retested delayed condition.

We also conducted the same gamma correlation analysis that Butler et al. (2011) had used to demonstrate the return of the errors. Here we use it to investigate the possibility (suggested by Fig. 4) that there might have been a selective return of high rather than low confidence errors in the untested delay condition, but not in the tested delayed condition. The analysis involved computing within-participant gamma correlations between confidence judgments for the original errors and ascription of correctness to these errors on a later test. As is shown in Fig. 5, the mean correlation was significantly greater than zero (indicating a return of the errors) in the untested delay condition ($M = .12$, $SD = .38$), $t(47) = 2.13$, $p = .04$, but not in the immediate test ($M = -.12$, $SD = .56$), $t(29) = 1.13$, $p = .27$, or in retested delay conditions ($M = -.07$, $SD = .51$), $t(45) = .90$, $p = .37$.[4] Thus, testing did prevent the return of the high confidence errors.

[4] Gamma correlations could not be computed for 19 participants in the immediate condition, 1 participant in the untested delay condition, and 3 participants in the retested delay condition.

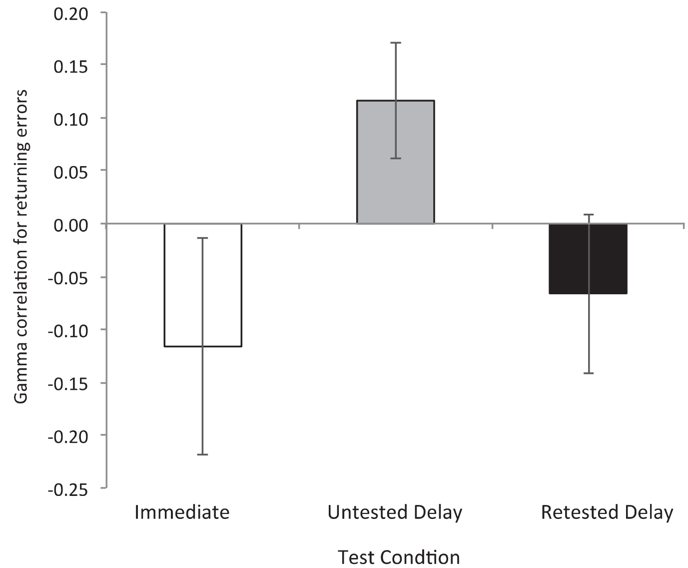

Fig. 5. Mean within-participant gamma correlations between confidence judgments for the original errors and ascription of correctness to these errors on a later test. Error bars reflect standard errors of the mean.

2.5.3. Ascription of correctness regardless of original confidence in errors

Collapsing across confidence and then computing means for each participant, there was a main effect of test condition for the number of errors that were later ascribed as being correct, $F(2, 96) = 53.77$, $p < .001$, $\eta_p^2 = .53$. The mean number of original errors, overall, that were later incorrectly ascribed as being correct was greater in the untested delay condition than in either the retested delay condition or the immediate condition ($ps < .001$). Although there was no main effect of regulatory focus on reoccurring errors, $F(1, 47) = 1.50$, $p = .23$, $\eta_p^2 = .03$, the regulatory focus × test condition interaction was significant, $F(2, 94) = 3.41$, $p = .04$, $\eta_p^2 = .07$, though none of the post hoc tests were significant. Participants were marginally more likely to endorse their original error on the delayed test of previously untested items the more prevention-focused they were ($r = -.26$, $p = .07$). This was not true for the immediate test or the delayed retest (rs ranged from $-.22$ to .13, $ps > .14$).

2.5.4. Did remembering the error help or hurt error correction?

This analysis is directed at the question of whether bringing to mind the original errors at time of test helped or hurt memory for the correct answers. It is relevant to the recursive reminding hypothesis (i.e., the possibility of a beneficial mediating effect of the errors on correct responding) on the one hand, and to the proactive interference hypothesis (i.e., the possibility that the mental presence of the errors results in a detrimental competition with the correct answers) on the other (see Butterfield & Metcalfe, 2001; Jacoby & Wahlheim, 2013; Wahlheim & Jacoby, 2013). We computed each participant's probability of giving the correct answer depending on whether the original error was present or absent as one of the two responses that were required on each question, and submitted these to a 2 (original error: present vs. absent) × 3 (test condition: immediate, untested delay, retested delay) repeated measures ANOVA. As is shown in the "Total" rows of Table 1, there was no main effect of the show up of the original error on correct responding, $F < 1$. The results did, of course, reveal the previously reported main effect of test condition, $F(2, 96) = 149.14$, $p < .001$, $\eta_p^2 = .76$, There was a marginally significant interaction between the show up of the original error × test condition, $F(2, 96) = 2.74$, $p = .07$, $\eta_p^2 = .05$. Uncorrected follow-up comparisons showed no

Table 1
Correct performance when original error was present or absent, by test condition and confidence in the original error. Performance could not be computed for some participants in certain instances due to a lack of qualifying trials (this is reflected in the *N* for each mean).

	Test condition								
	Immediate			Untested delay			Retested delay		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Original error *present*									
Low confidence (0–50)	48	72.40	25.46	49	30.90	21.65	48	46.15	24.97
High confidence (51–100)	42	79.64	29.43	46	43.54	33.52	42	64.81	34.73
Total (0–100)	49	74.04	23.89	49	36.24	20.37	49	51.33	22.66
Original error *absent*									
Low confidence (0–50)	49	69.33	17.07	49	35.49	19.33	49	56.57	19.61
High confidence (51–100)	36	76.78	30.14	32	51.28	35.32	40	72.97	32.02
Total (0–100)	49	71.08	17.54	49	37.27	18.13	49	58.57	16.77

difference depending upon whether the error showed up or not in either the immediate condition ($p = .46$), or in the untested delayed condition ($p = .73$), whereas the show up of the original error slightly hurt performance in the retested delayed condition ($p < .05$).

Next, we added the regulatory focus index to the previous ANOVA and conducted a repeated measures ANCOVA. The results revealed a marginally significant regulatory focus × test condition interaction, $F(2, 94) = 2.39$, $p < .10$, $\eta_p^2 = .05$, which represents the same interaction depicted in Fig. 1, but no main effect of regulatory focus, $F < 1$, no regulatory focus × original error interaction, $F < 1$, and no three-way interaction, $F(2, 94) = 1.13$, $p = .33$, $\eta_p^2 = .02$.

Finally, we also investigated whether the effects of bringing to mind the original error might have differed depending upon whether those errors had initially been made with high as compared to low confidence. According to the recursive reminding hypothesis, the show up of an error that was explicitly corrected on the pretest should help with retrieval of the correct answer on a later test. Although the previous analysis only showed an effect of error show up on performance in the retested delay condition (and this was in a direction inconsistent with the recursive reminding hypothesis), it is possible that the presence of the original error may have aided retrieval on trials in which the original error was made with high confidence. To test this possibility, we recomputed each participant's probability of giving the correct answer depending on the presence or the absence of the original error *and* confidence in the original error (errors that had a confidence judgment of over 50 on the slider scale were coded as high confidence errors, while errors that had a judgment of 50 or under were coded as low confidence error; see Table 1). We then submitted these to a series of six uncorrected paired sample *t*-tests that compared the presence versus absence of the original error for each combination of confidence (high vs. low) and test condition (immediate, untested delay, retested delay).[5] The only test that yielded a significant difference was for low confidence errors in the retested delay condition, $t(47) = 2.56$, $p = .01$ (for all other tests, $ts < 1.61$, $ps > .11$). Consistent with the previous analysis, the show up of original errors that were made with low confidence slightly hurt performance in the retested delayed condition. In all other cases, there was no relationship between error show up and correct performance.

[5] The reason we conducted separate analyses is because some participants lacked trials in which a high confidence error was either present or absent during a particular test. Conducting a repeated measures ANOVA across all three test conditions would have resulted in the exclusion of nearly half the sample. By conducting separate analyses for each combination of confidence and test condition, we were able to minimize exclusions.

3. Discussion

These results indicate that despite the strong hypercorrection effect shown by participants in all conditions, the high confidence errors did have a tendency to return at a delay when there was no intervening post feedback test. The most interesting result of this experiment – and one that has educational relevance – is that including a test immediately after the corrective feedback *protected against this return of the errors.*

This protection did not keep the errors from coming to mind. The rate of 'show up' of the original errors was the same in the retested and untested delayed conditions. But in the retested delayed condition, people subscribed to the errors as being correct to a lesser extent. The coming to mind of the errors in the retested delay condition slightly hurt correct performance: people were less likely to produce the correct answer and subscribe to it as correct when the original error was also produced at time of test. This finding is compatible with the interference theory notion of a competition between the original errors and the new response, or, alternatively that the erroneous responses need to be unlearned to allow new correct learning to occur. In the other conditions the mental presence of the errors was unrelated to whether the mistakes were or were not corrected.

The finding of a higher show up rate of high confidence errors as contrasted to low confidence errors as one of the two responses that participants were asked to generate at time of final test provides the basis for speculating that the recursive reminding hypothesis could potentially explain the hypercorrection effect. However, for recursive reminding to be an explanation of the hypercorrection effect, the bringing to mind of the errors would have to *facilitate* error correction. Unfortunately, as noted above, the show up of the original errors was either unrelated to whether the errors were corrected or negatively related. In no case was a positive relation observed.

While this result is disappointing, it does not necessarily falsify recursive reminding theory. A key idea in that theory is that the 'bringing to mind' of the original error is a marker that the entire *episode* in which the error was committed and corrected came to mind. Memory of this episode is proposed to include the question along with both erroneous and correct answers, as well as the memory of which answer was wrong and which right and perhaps other contextual information related to the event that occurred when the correction was provided. It is conceivable, though, that in the present experiment people's original errors might have come to mind semantically, and not be a signal that the correction event was retrieved. If the coming to mind of the errors did not indicate memory for the correction event, then the finding that the error show up was either unrelated or negatively related to correct responding does not falsify recursive reminding theory. The present results do limit the generality of the theory, though. It does not appear to be broadly applicable to semantic error correction, or to be a viable explanation of the hypercorrection effect.

Regardless of whether the errors were mentally present or not, people were less likely to ascribe to the original errors as being correct if they had previously had a test. Following a test, they were more likely to know that the correct answers were correct. Thus, having an immediate post-feedback test appears to have allowed the participants to encode the correct answers and to consolidate them *as* being correct.

The effects of regulatory focus on error correction, while interesting, were smaller than we had anticipated and provide no cause for concern to an educator. First, it was the error sensitive prevention-focused people in the present paradigm who, if anything, performed slightly *better*. The only condition in which this was significant, however, was in the untested delay condition – a condition in which memory for the correction was not protected by the intervention of an immediate test. The prevention-focused participants' concern with preventing mistakes may have motivated them to continue to think about their mistakes, as well as the corrections to their mistakes so as to not make them again. Finally, the prevention-focused participants tended to hypercorrect more in the immediate test condition than in the delayed test conditions. It is plausible that the attentional effects associated with hypercorrection (that may pertain particularly to prevention-focused individuals) are short lived. The semantic component of hypercorrection may be more important at a delay, and may be less vulnerable to regulatory focus effects. That the latent semantic analysis directed at evaluating the degree of associative relatedness of the low and high confidence errors to the target was the same for the prevention and promotion-focused people provides some credibility to this possibility.

3.1. Practical applications

From a practical standpoint, these results provide further support for the idea that having people generate errors (even errors in which they strongly believe) and then receive corrective feedback does not harm, and may substantially help, learning. It follows that instructors should not be concerned about the detrimental effects of encouraging students to generate errors as long as they provide students with corrective feedback. Indeed, high confidence errors were consistently hypercorrected in every experimental condition. The rate of correction, in all conditions, and for all participants, was impressive. This study, then, provides no grounds for evoking a caveat to the assertion that making errors, as long as those errors are followed by corrective feedback, helps students to learn.

Additionally, though, if nothing is done to prevent it, the high confidence errors return at a delay. Importantly, intervening a test immediately after corrective feedback thwarted the return of these errors. Thus, the intervention of a test after corrective feedback is highly advantageous. It appears to benefit the reconsolidation of semantic memory into a new corrected form that fortifies it against regression to earlier erroneous answers. The two practical recommendations for educators that follow from this research (but which require validation in classroom contexts), are (1) encourage students to generate their responses, even if they generate errors – but be sure to provide corrective feedback, and (2) consolidate the learning from that corrective feedback by administering an immediate test which will not only bolster recall but also prevent the return of the original errors.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

We thank the James S. McDonnell Foundation for supporting this research under Grant 220020166. We are also grateful for the help of Barbie Huelser, Joseph Bisaglio and Tory Higgins.

References

Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*, 97–105. http://dx.doi.org/10.1037/h0047507

Bohns, V. K., Lucas, G. M., Molden, D. C., Finkel, E. J., Coolsen, M. K., Kumashiro, M., et al. (2013). Opposites fit: Regulatory focus complementarity and relationship well-being. *Social Cognition*, *31*, 1–14.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin and Review*, *18*, 1238–1244.

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

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

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

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

Cesario, J., Grant, H., & Higgins, E. T. (2004). Regulatory fit and persuasion: Transfer from feeling right. *Journal of Personality and Social Psychology*, *86*, 388–404.

Crowe, E., & Higgins, E. T. (1997). Regulatory focus and strategic inclinations: Promotion and prevention in decision making. *Organizational Behavior and Human Decision Processes*, *69*, 117–132.

Eich, T. S., Stern, Y., & Metcalfe, J. (2013). The hypercorrection effect in younger and older adults. *Aging, Neuropsychology and Cognition*, *20*, 511–521.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

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. (2009, November). Persistence of the hypercorrection effect. *Poster presented at the annual meeting of the Psychonomic Society* Boston, MA,

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

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

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

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513.

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

Haws, K. L., Dholakia, U. M., & Bearden, W. O. (2010). An assessment of chronic regulatory focus measures. *Journal of Marketing Research*, *47*, 967–982.

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 290–296.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*, 1280–1300.

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., & Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, *31*, 3–23.

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, *32*, 336–350.

Hong, J., & Lee, A. Y. (2008). Be fit and be strong: Mastering self-regulation through regulatory fit. *Journal of Consumer Research*, *34*, 682–695.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning but learners do not know it. *Memory & Cognition*, *40*, 514–527.

Idson, L. C., & Higgins, E. T. (2000). How current feedback and chronic effectiveness influence motivation: Everything to gain versus everything to lose. *European Journal of Social Psychology*, *30*, 583–592.

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition*, *41*, 625–637.

Jacoby, L. L., Wahlheim, C. N., & Yonelinas, A. P. (2013). The role of detection and recollection of change in list discrimination. *Memory & Cognition*, *41*, 638–649.

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*, 989–998.

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

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

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

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

Metcalfe, J., Butterfield, B., Habeck, C., & Stern, Y. (2012). Neural correlates of people's hypercorrection of their false beliefs. *Journal of Cognitive Neuroscience*, *24*, 1571–1583.

Metcalfe, J., & Finn, B. (2012a). People's hypercorrection of high confidence errors: Did they know it all along? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 437–448.

Metcalfe, J., & Finn, B. (2012b). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*, 253–261.

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

Molden, D. C., & Miele, D. B. (2008). The origins and influences of promotion-focused and prevention-focused achievement motivations. In M. Maehr, S. Karabenick, & T. Urdan (Eds.), *Advances in motivation and achievement: Social psychological perspectives on motivation and achievement* (15) (pp. 81–118). Bingley, UK: Emerald.

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

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

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3–8) http://dx.doi.org/10.1037/0278-7393.31.1

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

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*, 243–257.

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Scholer, A. A., Ozaki, Y., & Higgins, E. T. (2014). Inflating and deflating the self: Sustaining motivational concerns through self-evaluation. *Journal of Experimental Social Psychology*, *51*, 60–73.

Scholer, A. A., Stroessner, S. J., & Higgins, E. T. (2008). Responding to negativity: How a risky tactic can serve a vigilant strategy. *Journal of Experimental Social Psychology*, *44*, 767–774.

Shu, T.-M., & Lam, S. (2011). Are success and failure experiences equally motivational? An investigation of regulatory focus and feedback. *Learning and Individual Differences*, *21*, 724–727.

Sitzman, D. M., & Rhodes, M. G. (2010). Does the hypercorrection effect occur when feedback is delayed? *Poster Session Presented at the Psychonomics Society Annual Meeting, #5090*.

Sitzman, D. M., Rhodes, M. G., & Tauber, S. K. (2014). Prior knowledge is more predictive of error correction than subjective confidence. *Memory & Cognition*, *42*, 84–96.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, *22*, 153–163.

Van-Dijk, D., & Kluger, A. N. (2004). Feedback sign effect on motivation: Is it moderated by regulatory focus? *Applied Psychology: An International Review*, *53*, 113–135.

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive reminding in proactive effects of memory. *Memory & Cognition*, *41*, 1–15.