Older adults fail to show a hypercorrection effect

Teal S. Eich¹, Yaakov Stern¹,² and Janet Metcalfe¹

¹Department of Psychology, Columbia University, New York, NY, USA
²Cognitive Neuroscience Division, Taub Institute for Research in Alzheimer’s Disease and the Aging Brain, College of Physicians and Surgeons of Columbia University, New York, NY, USA

Accepted (with revision) Psychology and Aging, June 14, 2011
Not yet final. Please do not quote (yet! ; ).
Abstract

The hypercorrection effect refers to the finding that errors that are committed with high confidence are more likely to be corrected than are low confidence errors. We investigated whether older adults, like younger adults, show a hypercorrection effect. Older and younger adults answered general-information questions, made confidence ratings about their answers, were given corrective feedback, and then were retested on questions that they had gotten wrong. Younger adults showed the hypercorrection effect, as has been shown previously. Older adults, despite higher overall accuracy on the general-information questions and excellent basic metacognitive ability, however, failed to show the hypercorrection effect.

Keywords: Hypercorrection Effect; Metacognition; Older Adults; Confidence; Updating.
People's ability to correct their errors is central for learning in all domains. Much research indicates that corrective feedback is necessary for this updating process (Pashler, Cepeda, Wixted, & Rohrer, 2005). Many studies, however, have demonstrated that the effectiveness of corrective feedback is modulated by how confident people are in their errors. While it seems intuitive that correcting errors committed with high confidence should be more difficult than correcting less entrenched, low confidence errors, the data on the relation between confidence in one’s error and the probability of subsequent recall of the correct answer show the opposite result: correct recall performance is better for errors that were initially committed with high, rather than low confidence. This finding is called the hypercorrection effect (Butterfield & Metcalfe, 2001).

In the hypercorrection paradigm, participants answer general-information questions and then rate how confident they are in their answers. If they answered incorrectly, they are immediately provided with corrective feedback. After reaching a criterion number of incorrect responses, the participants are re-tested on the questions they got wrong. Butterfield and Metcalfe (2001) found that the correct answer was more likely to be given on the final test for questions that the participants had originally made a high confidence error, as opposed to an error made with low confidence.

This effect has been replicated numerous times, and persists across multiple variants of the original task. Butterfield and Metcalfe (2006) replicated the hypercorrection effect in a task that included a tone detection element. Metcalfe, Butterfield, Habeck, and Stern (under review) replicated it in an imaging (fMRI) study. Fazio and Marsh (2009) replicated the hypercorrection effect, and also found that participants were more likely to remember surface features (the color
of the text or the sex of the feedback-voice) presented in conjunction with feedback to high, but not low confidence errors. Metcalfe and Finn (2011) replicated the basic findings using four different task variations, and showed that people frequently claimed that they “knew it all along” when they heard the answer to high (but not low) confidence errors. Kulhavy and Stock’s (1989) study demonstrated a hypercorrection effect that they attributed to superior comprehension of high confidence errors. Butterfield and Mangel’s (2003) study demonstrated that hypercorrection persists even when initial errors were tested one week later. Butler, Fazio and Marsh (2010), similarly, used the standard procedure for testing hypercorrection, and re-tested incorrect answers both immediately and after a 1-week delay. They replicated the effect, and showed that it persisted. Sitzman and Rhodes (2010) also found that re-test accuracy for initially incorrect responses was greater for high confidence, as opposed to low-confidence errors after a 2-day delay. Butler and Roediger (2008) found the effect using a multiple-choice paradigm. Hypercorrection has even been shown to occur for false memories (Fazio & Marsh, 2010). Finally, Metcalfe and Finn (under review), in four separate experiments, observed the hypercorrection effect with 8 to 12 year old children. The hypercorrection effect, then, appears to be robust, and is shown by both children and young adults. The question we address here is whether the hypercorrection effect holds up throughout the entire lifespan. In particular: Do older adults show it?

There have been several analyses of the mechanisms underlying error correction, in general, and the hypercorrection effect, in particular. While some of these suggest that older adults may show the effect, others suggest that they might not. One observation related to the hypercorrection effect is that both the target items, themselves, and the domains of the questions are more familiar and semantically rich for high than for low confidence errors (Butterfield &
Metcalfe, 2006; Butterfield & Mangels, 2006; Metcalfe & Finn, 2011). When people feel that they have more knowledge about the domain, their confidence in their response is higher (see Metcalfe, Schwartz & Joaquim, 1993; Schwartz & Metcalfe, 1992; Metcalfe & Finn, 2008). While the correct response to a high confidence error has been shown to be familiar to the participant, the correct response to a low confidence error does not elicit much if any familiarity (Metcalfe & Finn, 2011).

To the extent that familiarity contributes to learning, and the familiarity is greater for the high than for the low confidence targets (and contributes to the hypercorrection effect), we would expect older adults to show the effect. Brainerd, Reyna and Howe (2009) conducted careful parameter estimates using their fuzzy trace model to isolate the components of remembering with younger and older adults. They concluded that the familiarity parameter is intact with aging. Reviews of older adults’ cognitive capabilities have shown that both general knowledge and semantic memory (Verhaeghen, 2003) are spared. Similarly, if the hypercorrection effect were attributable to participants’ metacognitive ability, it is also likely that older adults should show the effect, since, typically, metacognition is spared in healthy older adults (Hertzog, 2002; Allen-Burge & Storandt, 2000).

Older adults have been shown to exhibit difficulty with new learning (Gilbert, 1941; MacKay & Burke, 1990; Hedden & Gabrieli, 2004). Brainerd, Reyna and Howe (2009) showed that older adults have a pronounced deficit in 'access' memory—or what, in the error correction paradigm, could be the memory for the just-presented specifics of the correct answer. To the extent that correcting their initial responses entails new learning of a new response, it is possible that, because of this difficulty with access or new specific learning, the older adults might experience greater difficulty. However, it is not obvious why the failure of access-memory
would necessarily affect the learning of new responses for *high* confidence errors selectively. It would seem that although this impairment might predict an impairment in overall error correction, it should occur equally for high and low confidence errors.

There are two factors, however, that might result in a *selective* hypercorrection failure in older adults. First, the error-correction task requires that the old, incorrect, responses, be suppressed. This need to suppress would seem to be selective to high confidence errors, since these are strongest (and presumably most intrusive) errors. We have recently (Metcalfe, Butterfield, Habeck & Stern, submitted) shown that the same brain area that is involved in the suppression of unwanted memories in the think/no think paradigm (i.e., right dorsolateral prefrontal cortex, as has been demonstrated by Anderson, Ochsner, Kuhl, Cooper, et al, 2004) is also activated when college-aged students receive feedback to high (but not low) confidence errors. It is of considerable interest that Anderson, Reinholf, Kuhl, and Mayr (2011) have shown that older adults have particular difficulty in suppressing unwanted memories in the think/no think paradigm. Older adults’ difficulty in other kinds of suppression has been shown by others as well (Gazzaley, Clapp, Kelley, McEvoy, Knight & D'Esposito, 2008; Hasher, Zacks and May, 1999; Hasher, Chung, May & Foong, 2002; Healey, Campbell, Hasher & Ossher, 2010; Jacoby, Bishara, Hessels & Toth, 2005). Thus, to the extent that suppression may be necessary for hypercorrection to occur, a selective impairment in older adults is predicted.

Finally, it has been proposed that enhanced encoding of the high confidence errors arises because of the metacognitive mismatch between the person’s expectations about the correctness of their response and the fact that they are wrong. This explanation of the effect (Butterfield & Metcalfe, 2006; Butterfield & Mangels, 2003; Fazio & Marsh, 2010) proposes that people attend more to the feedback to high as compared to low confidence errors because participants’ surprise
at being wrong, when they were very sure they were right, is a negative outcome that increases attention to the correct answer, which in turn facilitates memory. In Butterfield and Metcalfe’s (2006) study a simultaneous tone-detection task was presented during the corrective feedback in an effort to assess this attentional-capture hypothesis. Tone detection was selectively impaired during high confidence error feedback, in support of the hypothesis. Fazio and Marsh's (2009) data, indicating enhanced contextual memory for the metacognitive mismatch conditions, also supports an attention explanation. Finally, Butterfield and Mangels (2003) found evidence consistent with the attentional-capture hypothesis. Using event related potentials (ERPs), they found a confidence-modulated p300 event related potential deflection-- similar to the novelty-P3/P3a, often associated with increased surprise-related attention--that occurred to the corrective feedback in a graded way that increased systematically with higher confidence in the errors.

Consistent with these ERP findings, Metcalfe, Butterfield, Habeck and Stern (submitted) found anterior cingulate activations associated with the feedback to high confidence error feedback, suggesting that high confidence errors were accompanied by conflict and negative emotions.

Older adults, including high functioning, cognitively intact participants, sometimes show a deficit in surprise-related attentional processing in other paradigms. They sometimes (but not always, see Emery, Hale & Meyerson, 2008) show differences in release from proactive inhibition (Hasher, Chung, May & Foong, 2002). They often show a failure to benefit memorialy from a novelty item in a von Restorff isolation situation (Cimbalo & Brink, 1982), although sometimes the effects are small (e.g., Bireta, Surpreant & Neath, 2008). Older adults show impaired P3 ERP deflections, both in terms of their latency and amplitude, in standard oddball tasks (see Fjell & Walhovd, 2004). They also exhibit a positivity bias, whereby they tend to be less responsive to negative emotional outcomes than do younger adults (Wood &
Kisley, 2006; Grühn, Scheibe & Baltes, 2007). Thus, if the increased attention due to affectively negative metacognitive mismatch is critical for hypercorrection to occur, older adults might not show the hypercorrection effect.

One final point: there are many studies showing that older adults are slower than younger adults (see Der & Deary, 2006)—a factor that might have an effect in our paradigm, if, for example, it simply took more time for the older adults to fully process the questions and make their answers. We did not want penalize the older adults for nonspecific differences such as their speed of responding. Therefore, in the present experiment, we employed two slightly different versions of the standard hypercorrection task, with the amount of time for which the correct response was displayed (either 1.5 or 4 sec) varying as a between-participants manipulation. Generally, we thought that the older adults might not show the hypercorrection effect at fast rates--as they might not have sufficient time to encode the correct answer--but that they might show the effect at the slower rate of feedback presentation.

**Method**

**Participants.** The participants were 51 (40F/11M, average age 20.7) Columbia University and Barnard College students, who participated for course credit, and 31 older adults (17F/14M, average age 65.7), recruited through fliers, internet advertisement, market mailings and from senior day centers located in Manhattan, New York. The older adults were paid for their participation. The older adults were screened to ensure the absence of any past or current medical, neurological, or psychiatric disorders, including dementia, or treatment with psychoactive drugs. All participants were treated in accordance with APA ethical guidelines.

**Materials.** Participants were asked general-information questions taken in part from the set of Nelson and Narens (1980) and augmented as described by Metcalfe and Finn (2011).
Examples of questions were, “What is the river than runs through Rome?” (answer: Tiber) or “Which actor thanked his parents for not using birth control upon receiving a 1979 Oscar?” (answer: Hoffman).

**Procedure.** Participants, tested individually, were presented, one at a time, with general-information questions. A question appeared in the center of the computer screen, and the experimenter read it aloud. The participant was instructed to answer or give his or her best guess to the question, and the response was recorded by the experimenter. Once he or she had responded, the participant made a confidence rating about his or her answer using a horizontal slider on the computer that ranged from “very unsure” on the left end to “very sure” on the right end. Confidence ratings were coded along a scale from 0 to 1, with 0 indicating a selection of the lowest limit of the slider, at the very unsure end, and 1 indicating a selection of the highest limit, at the very sure end. The slider bar was anchored to the middle of the scale at the onset of each question, so that moving up was in the direction of high confidence while moving down was in the direction of low confidence. The participant had to point to the chosen location on the scale, and the experimenter then moved the cursor to this position. Once a confidence rating was made, the participant received feedback. When the participant’s answer was correct, a chime sounded and the next question was presented immediately. When the answer was incorrect, the correct answer was presented on the screen and the experimenter read it aloud. Once 15 incorrect items had been accumulated, the program randomized those 15 originally incorrect responses, and retested each, for a final cued-recall test. Just as in the original test, the question appeared in the center of the computer screen, the experimenter read it aloud and then entered the participant’s response. At the end of the experiment, all participants were thanked for their participation and debriefed.
**Design.** We employed a 2 X 2 between participant factorial design, with Age (young or old) and the amount of Feedback Time (1.5s or 4s) manipulated.

**Results**

**Basic data.** Older adults were better at correctly answering the general-information questions than were the younger adults. As can be seen in Figure 1, the number of questions that had to be presented in order to reach 15 erroneous responses was significantly greater for the older as compared to the younger adults, (\( \bar{X}=22.03, SD=4.96 \) vs. \( \bar{X}=19.33, SD=2.99 \) respectively; \( F(1,78)=8.81, MSE=15.13, p=.004 \)). Neither the main effect of Feedback Time, nor and the interaction between Age and Feedback Time were significant (both \( F's < 1 \)). Older and younger adults differed in their confidence ratings. The older adults’ mean confidence ratings for errors in the 1.5s and 4s conditions were \( \bar{X}=34.59 \) (\( SD=19.16 \)) and \( \bar{X}=25.98 \) (\( SD=13.66 \)); for younger adults they were \( \bar{X}=16.36 \) and \( \bar{X}=15.05 \) (\( SD=8.41 \) and 9.33). There was a main effect of Age, such that older adults made significantly higher confidence ratings than did younger adults (\( F(1,78)=27.39, p=.000 \)). Neither the main effect of Feedback Time, nor and the interaction between Age and Feedback Time were significant (\( F(1,78)=3.18, p=.08; F(1,78)=1.72, p=.19 \) respectively). We also computed the average confidence rating for young and old for incorrect and for correct responses on the first test separately. For correct responses, older adults averaged .80 confidence on a scale from 0-1, while younger adults averaged .67. For incorrect responses, older adults averaged .30 confidence, while younger adults averaged .16.

To investigate whether older and younger adults’ metacognitive abilities were comparable (and that confidence ratings were predictive of initial test performance), we computed the gamma correlations between confidence ratings and accuracy on the first test.
The gamma statistic tests the strength of association between cross-tabulated data. Values range from $-1$ (100% negative association, or perfect inversion) to $+1$ (100% positive association, or perfect agreement). A value of zero indicates the absence of association. (In subsequent analyses we were sometimes unable to report a gamma correlation for some participants because some got everything right or everything wrong, or had too many ties and the statistic could not be computed. Thus, degrees of freedom listed for gamma correlations may differ from the total number of participants used in the experiment). Mean gamma values were high in both the 1.5s and 4s conditions for the older adults ($\gamma = .85, SD = .15$ and $\gamma = .81, SD = .16$), and for the younger adults ($\gamma = .79, SD = .16$ and $\gamma = .79, SD = .24$), and all four gammas were significantly greater than zero ($t(12)=20.14; t(16)=20.90; t(22)=23.52; t(25)=16.51; all p<.001$).

No significant differences were found among conditions, however ($Fu<1$). These results indicate that the older adults had excellent basic metacognition on this task: they knew what they knew and what they did not know, and were every bit as good, metacognitively, as the younger adults. Mean, overall, post feedback recall performance was the same for older and younger adults ($F<1$) (see Figure 1). The older adults’ mean recall in the 1.5s and 4s conditions were $\bar{X}=.76 (SD=.23)$ and $\bar{X}=.80 (SD=.16)$; for younger adults they were $\bar{X}=.77$ and $\bar{X}=.75 (SD=.12$ and .14).

**The Hypercorrection Effect.** A hypercorrection effect would be in evidence if high confidence errors were more likely to be corrected on the final test than errors endorsed with lower confidence. Accordingly the gamma correlations between confidence rating on the original errors and accuracy on the final test was the dependent variable of interest, with positive values indicating hypercorrection. Accuracy on the final test was scored leniently by an algorithm created by Butterfield based on letter overlap that allowed as correct what most
human scorers would consider to be only spelling mistakes (e.g., Hofman for Hoffman). The data were also hand checked by the experimenters, to ensure that no responses that could have been simply spelling mistakes were considered to be errors. As can be seen in Figure 1, while the younger adults hypercorrected (γ = .51, SD = .55), the older adults did not (γ = .14, SD = .68). As well as being different from one another (F(1, 63) = 5.87, p = .02), one-sample t-tests revealed that while younger adults’ gammas were significantly greater than zero (t(41) = 5.99, p = < .001), those of the older adults were not (t(24) = 1.05, ns). Neither the main effect of Feedback Time (F(1, 63) = 2.68, ns) nor the interaction between Age and Feedback Time (F(1, 63) = 1.36, ns) were significant.

To investigate the hypercorrection effect, in addition to the gamma statistic we also split the data into the highest 7 and lowest 7 confidence errors for each participant (omitting the response in the middle, if it existed) and computed the conditional probabilities over these high and low confidence responses, for each subject. The probability of error correction for the young subjects was .76 and .85, for low and high confidence errors (t(1, 50) = 2.954, p = .005). This analysis, while less sensitive that the gamma correlation, showed that the younger participants hypercorrected. The elders' probability correct were .81 and .84 for low and high confidence errors (t(1, 30) = 1.07, p = .292), a result that substantiated the conclusion from the gamma correlation, that they failed to hypercorrect.

**General Discussion**

The results reported here for the younger adults replicated numerous previous studies’ findings that confidence in one's errors impacts which errors are most likely to be corrected: errors that were endorsed with high confidence were hypercorrected following corrective
feedback, while those made with low confidence were not. However, we found no such effect with the older adults. They did not show the now-typical hypercorrection effect.

Although reviews of older adults’ cognitive capabilities have shown that both general knowledge and semantic memory (Verhaeghen, 2003) are spared, to assess whether the familiarity of the materials might have been different between participant groups, and whether this may have contributed to our finding of a lack of a hypercorrection effect in older adults, we performed the same analysis Metcalfe and Finn (2011) had used to compare the semantic association strength between the target word (e.g., “Ottawa,” to the question: “What is the capital of Canada?”) and the error that the participant made (e.g., “Toronto”) using the Latent Semantic Analysis (LSA) tool. LSA provides a quantitative measure of the degree to which two word senses are related (Landauer & Dumais, 1997, and see http://cwl-projects.cogsci.rpi.edu/msr/).

The LSA values for low confidence errors were .26 ($SD=.08$) for the younger adults and .20 ($SD=.1$) for the older adults. For high confidence errors, the LSA values respectively for younger and older adults were .36 ($SD=.19$) and .34 ($SD=.16$). An ANOVA with LSA value as the dependent variable, and contrasting Confidence (high or low), Age (young or old) and Feedback time (1.5 or 4 s) revealed a significant effect only for high versus low confidence responses ($F(1,60)=24.58, p<.001$). No other effects or interactions were significant, and, in particular, there were no effects or interactions with age. Thus, it does not appear to be the case that the familiarity differences could explain the lack of a hypercorrection effect in the older adults.

Although we do not dispute that older adults may have an access deficit, it was not manifested in the present general-information paradigm. If there had been such a new learning or access-memory deficit in this experiment, it should have been evident as a main effect on all
error correction. However, the older adults did very well on the task, and corrected their errors, overall, slightly more so than did the younger adults. Furthermore, the errors that they had to correct were likely to have been, on balance, slightly more difficult than those of the younger adults, insofar as the older adults committed fewer, not more, errors on the first phase of the experiment. The lack of an overall access-based impairment suggests that, at least in the present domain of general-information fact learning, poor access was unlikely to have been a factor that impacted the older adults' performance.

There remain (at least) two viable hypotheses for why the older adults failed to hypercorrect. The lack of a hypercorrection effect might have resulted because the older adults had difficulty suppressing the incorrect alternatives during feedback sufficiently to learn the new responses. While the data did not show a large number of intrusions of the original erroneous responses for either group (10 out of 460 total trials for older adults, 3 out of 753 total trials for younger adults), the increased effort needed for the older adults to suppress the wrong but highly confident responses might have interfered with their learning of the correct responses.

Alternatively, the metacognitive-mismatch hypothesis says that people rally attention when they realize that they have committed a high confidence error. Older adults may have failed to do so to the same extent as younger adults, because they were either not as surprised as younger adults at being wrong, or because they maintained a positivity bias (Charles, Mather & Carstensen, 2003; Mikels, Larkin, Reuter-Lorenz & Carstensen, 2005), or at least a lack of the negativity bias (see Ito, Larsen, Smith, & Cacioppo, 1998). It is possible that younger adults find high confidence errors to be negative events that are aversive and possibly embarrassing, but that older adults—shunning such a negativity bias—did not find them so emotional and hence did not selectively attend to or correct them. At the present time, we cannot tell which of these
hypotheses (or perhaps, others, that we have not considered) were responsible for the selective failure to hypercorrect. Further investigation into the possible mechanisms underlying this effect is needed.

In conclusion, we demonstrate for the first time that, unlike younger adults and children, older adults fail to hypercorrect. Older adults did not show a deficit in general-information knowledge. They were also not impaired in their metacognitive assessments. But, the pattern of their error correction was quite different from that of younger adults: the older adults, unlike younger adults and even children, did not selectively correct their high confidence errors.
Figure 1: Performance measures for older and younger adults in the hypercorrection paradigm:
Accuracy for first and final tests, and Gamma correlations between first test confidence ratings and final test accuracy.
Acknowledgements

This research was supported by Grant 220020166 from the James S. McDonnell Foundation. The authors are wholly responsible for the content of this article. We would like to thank Philip Yeung for his assistance in recruiting and scheduling older adults, David Miele and Brandon Luke for their help testing younger adults, and Barbie Huelser for her help with the LSA analysis.
References


