Beyond Hypercorrection:

Remembering Corrective Feedback for Low-Confidence Errors

Lauren Griffiths and Philip A. Higham

University of Southampton

Author Notes

 Lauren Griffiths, Department of Psychology, University of Southampton; Philip A. Higham, Department of Psychology, University of Southampton.

 Portions of this research were presented at 56th Annual Meeting of the Psychonomic Society, Chicago, USA, November, 2015.

Correspondence concerning this article should be addressed to Philip A. Higham, Department of Psychology, University of Southampton, Highfield, Southampton, UK, SO17 1BJ. Tel: +44 (0) 23 8059 6652. E-mail: P.A.Higham@soton.ac.uk.

Abstract

Correcting errors based on corrective feedback is essential to successful learning. Previous studies have found that corrections to high-confidence errors are better remembered than low-confidence errors (the hypercorrection effect). The aim of this study was to investigate whether corrections to low-confidence errors can also be successfully retained in some cases. Participants completed an initial multiple-choice test consisting of control, trick and easy general-knowledge questions, rated their confidence after answering each question, and then received immediate corrective feedback. After a short delay, they were given a cued-recall test consisting of the same questions. In two experiments, we found high-confidence errors to control questions were better corrected on the second test compared to low-confidence errors – the typical hypercorrection effect. However, low-confidence errors to trick questions were just as likely to be corrected as high-confidence errors. Most surprisingly, we found that memory for the feedback and original responses, not confidence or surprise, were significant predictors of error correction. We conclude that for some types of material, there is an effortful process of elaboration and problem solving prior to making low-confidence errors that facilitates memory of corrective feedback.

*Keywords*: hypercorrection, error correction, retrieval practice, confidence, corrective feedback

Beyond Hypercorrection:

Remembering Corrective Feedback for Low-Confidence Errors

Consider a student who chooses to answer a question that her instructor has posed during a lecture, but she gets the answer wrong. Her instructor politely corrects her and continues with the lecture. Will this student learn from this experience? Will the corrective feedback provided by her instructor be retained in her memory so that when she encounters the same question again on a final exam, she corrects her previous error? What if she answered the question correctly? Would the confirmatory feedback have any effect on her later test performance?

Both corrective and confirmatory feedback is ubiquitous in education and takes many forms. Perhaps for this reason, the role of immediate feedback on learning in educational contexts has been a topic of research in education and psychology for decades (e.g., see Kluger & DeNisi, 1996; Kulhavy & Stock, 1989; Kulik & Kulik, 1988; Nicol & MacFarlane-Dick, 2006 for reviews). One strand of this research has focused on memory for the feedback. After all, students who cannot remember feedback on a final exam are perhaps no better off than students who received no feedback at all. This research on feedback memory has highlighted the central importance of learners’ subjective confidence in their initial, pre-feedback response.

 To illustrate this importance, consider the four potential outcomes depicted in Figure 1. The figure presents a scenario in which learners respond to test questions on an initial test (Test 1) and then receive corrective feedback immediately following each response. Learners then take a second test after a delay that contains some or all of the questions that appeared on Test 1.[[1]](#footnote-1) What impact, if any, will the feedback on Test-1 responses have on Test-2 performance? Consider first the path leading to Outcome A at the top of the figure. In this scenario, learners made a correct Test-1 response with high confidence and the feedback confirmed that the response was correct. Under these circumstances, the feedback is usually superfluous because, with or without it, the correct answer is often firmly entrenched in memory and likely to be (re)produced on any test (e.g., Butler, Karpicke, & Roediger, 2008; Guthrie, 1971; Kulhavy & Anderson, 1972; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler, Rohrer, Cepeda, & Carpenter, 2007).

Consider now Outcome B for which a correct answer is produced with low confidence (e.g., a lucky guess on a multiple-choice exam). Unlike the first case, the feedback in the Outcome B path creates a *metacognitive mismatch* (Butterfield & Mangels, 2003); that is, the low confidence associated with the answer mismatches its correctness. The mismatch means that the feedback is surprising, which, in turn, causes it to attract attention and to be processed more deeply to resolve the discrepancy. The result is that low-confidence correct responses are more likely to be reproduced on Test 2 if feedback is provided than if it is not. As evidence for this finding, Butler et al. (2008, Experiment 2) investigated the effect of feedback on memory for correct responses across a two-day retention interval. They found that the benefit of feedback on final-test performance was greatest for low-confidence correct responses on the initial test (i.e., when participants were essentially guessing). Butler et al. concluded that feedback serves to correct metacognitive mismatches between perceived and actual accuracy (see also Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009).

Another type of metacognitive mismatch occurs if incorrect responses are produced on Test 1 with high confidence, which results in Outcome C in Figure 1. For many years, the consensus amongst learning theorists was that errors should be avoided at all costs during the learning process because of proactive interference. On this view, when an error is corrected with feedback, the original error competes with the newly acquired correct information. Furthermore, because the misinformation already has enough memory strength to be retrieved in response to a question, and has gained additional strength from the very process of being retrieved, interference theory (e.g., Muller & Schumann, 1894; Webb, 1917) predicts that the corrective feedback will likely be forgotten and the original error will reoccur in the future (see also Anderson & Reder, 1999; Barnes & Underwood, 1959; Melton & Irwin, 1940; Osgood, 1949). This interference is likely to be particularly great if errors are produced with high confidence because such errors are likely to also have high memory strength, making them unlikely to be forgotten.

However, contrary to these predictions of interference theory, in recent years much research has shown that correcting high-confidence errors with feedback can in fact aid learning (e.g. Butler et al., 2008; Butler & Roediger, 2008; Finn & Metcalfe, 2010; Guthrie, 1971; Kang, McDermott & Roediger, 2007; Kulhavy, Yekovich, & Dyer, 1976; Mullet, Butler, Verdin, von Borries & Marsh, 2014, although see Butler et al., 2011 who showed that after a one-week delay, some original high-confidence errors can return). For example, Butterfield and Metcalfe (2001) asked participants to answer general-knowledge questions and rate their confidence in their responses until they produced 15 correct and 15 incorrect answers. Corrective feedback was given immediately after each item. After a five-minute filler task, participants were given a final retest for which they were required to write down three possible responses to each question before indicating with a star the response they believed to be correct. They also rated their confidence for each final (starred) response. Butterfield and Metcalfe found that high-confidence errors were more likely to be present in the list of three possible responses at retest than low-confidence errors. However, contrary to the prediction based on interference theory, those responses were seldom starred. Instead, participants starred the correct answers (i.e., the feedback) such that they were actually more likely to correct high-confidence errors than low-confidence errors on the final test, an effect Butterfield and Metcalfe termed *hypercorrection*. Similar to the explanation for why feedback for low-confidence correct responses are remembered, Butterfield and Mangels (2003) argued that participants became more aroused due to the metacognitive mismatch between their high confidence and the fact that the response was an error. This increased arousal led them to pay more attention to the feedback to resolve the discrepancy, which facilitated encoding and storage of the correct answer for future use.

Follow up research on the hypercorrection effect has supported Butterfield and Mangels’ (2003) arousal account. For example, using a tone detection task, Butterfield and Metcalfe (2006) found that participants were able to detect the tone when it accompanied feedback for low-confidence errors but were less able to detect the tone when attending to feedback for high-confidence errors. A significant inverse relationship between tone detection and final performance was also found, which suggests that participants failed to detect the tone on high-confidence error trials because they were encoding corrective feedback. In a similar vein, Fazio and Marsh (2009) found that source information (Experiment 1: feedback colour; Experiment 2: gender of speaker who presented the feedback) was more likely to be remembered if the feedback was incongruent with participants’ expectations (i.e., there was a metacognitive mismatch: low-confidence correct responses and high-confidence errors).

Prior knowledge may also play a key role in the hypercorrection effect. Metcalfe and Finn (2011) found that participants were more likely to say that they “knew all along” the corrective feedback to questions they initially answered incorrectly with high versus low confidence. Furthermore, when given the opportunity to make a second response after making an error (with no feedback), participants were more likely to (a) select the correct answer from a list of options and (2) generate the correct answer, if the initial error was made with high- compared to low-confidence. These finding suggest that there is some validity to participants’ claims that they “knew all along” the correct answers to questions initially answered incorrectly with high confidence. Sitzman, Rhodes, and Tauber (2014) and Sitzman, Rhodes, Tauber, and Licerdale (2015) also found a positive relationship between the level of prior domain knowledge for a question and the likelihood of correcting an error. They found that when participants had little or no prior knowledge of a topic, they were unlikely to correct their error despite receiving feedback. They suggested that confidence ratings merely serve as a proxy for prior knowledge. Therefore, it is not the discrepancy between *confidence* and feedback that produces the hypercorrection effect, but a discrepancy between *prior knowledge* and feedback. Importantly, regardless of whether it is high confidence per se or high levels of prior knowledge that is the cause of hypercorrection, the outcome is largely the same: participants pay close attention to the feedback, they encode it efficiently, and they retain it well in memory (i.e., Outcome C in Figure 1).

Finally, consider Outcome D in Figure 1, the case of low-confidence errors. This particular scenario has been largely ignored in the literature because such cases appear to lack the ingredients (such as a metacognitive mismatch) needed for feedback to facilitate error correction. As Kulhavy et al. (1976) argued, “people make low-confidence responses primarily because they have difficulty understanding the material, the question, or both…If a student is having trouble understanding what he reads, providing feedback after he guesses at an answer should do little to improve comprehension” (p. 524).[[2]](#footnote-2)

Indeed, some research has shown that low-confidence errors are more likely to repeatedly intrude on subsequent tests if they were initially produced due to lack of understanding (e.g., Elley, 1966; Kulhavy & Parsons, 1972). Participants may also produce low-confidence errors as a result of inattention or a lack of prior knowledge about the domain of the question. In cases such as these, it may seem intuitive that feedback will do little to improve the situation. The primary focus of the current research is to question these assumptions about remembering feedback and correcting low-confidence errors.

**Current Study**

Although situations like those described above that lead to Outcome D are unlikely to lead to good error correction, there may be other situations in which low-confidence errors are retained well. For example, a student attempting to answer a multiple-choice question may identify two options that seem equally plausible and subsequently spend considerable effort elaborating on the relevant material in an attempt to select between them. Conceivably, if an option is eventually chosen and corrective feedback indicates that it was incorrect, that feedback may be retained quite well, perhaps as well as in cases of metacognitive mismatches (Outcomes B and C). To simulate this type of scenario, in two experiments, we presented general-knowledge questions with four multiple-choice alternatives to participants in a lab-based computer task, provided corrective feedback for each response, and then retested the same questions in a final cued-recall test. We used two types of critical questions, control and trick, designed to induce both high- and low-confidence errors. We also included some easy filler questions so that the test was not too difficult overall.

We hypothesized that multiple-choice “trick” questions could produce two patterns of responding. First, if the trick is missed, participants are likely to choose the easy, incorrect option with high confidence. For example, the question *“How many of each type of animal did Moses take onto the Ark? (a) 0 (b) 1 (c) 2 (d) 4”* is likely to induce the error *“(c) 2”* instead of the correct answer *“(a) 0”* because participants fail to notice that the question refers to Moses instead of Noah. Erikson and Mattson (1981) termed this phenomenon the *Moses Illusion* and suggested that this type of trick question occurs when it contains superfluous misinformation (e.g., names could be removed altogether from the above example: *“How many of each type of animal were taken onto the Ark?”*) and the misinformation within the question is semantically related to its true counterpart (i.e., *Moses* is related to *Noah*). After receiving corrective feedback to this type of trick question, participants are likely to be surprised (i.e., experience a metacognitive mismatch; Butterfield & Mangels, 2003) and show good memory for the feedback. This result, of course, is the typical finding found in the hypercorrection literature.

However, other trick questions might cause participants to engage in elaborative retrieval processes or problem solving prior to answering the question and receiving feedback, a situation more akin to the hypothetical student alluded to above who was struggling to discriminate between two equally-plausible options to a multiple-choice question. We hypothesise that these elaborative processes, in turn, could effectively potentiate learning of the corrective feedback by increasing the amount of attention devoted to processing the feedback, which leads to enduring memories (Richland, Kornell, & Kao, 2009). Critically, we predict this outcome even though confidence in the initial incorrect answer was low. In particular, we suspect that it will be the combination of competing multiple response options to choose from combined with one option that is “too easy” that will lead to this type of potentiation.

As an example of the second case, consider the trick question *“If beer is produced in a brewery, what is produced in a ginnery? (a) Cotton (b) Gin (c) Wool (d) Vodka”*. Without knowing the correct answer, which is *“(a) Cotton”*, the obvious, intuitive answer to this question would be *“(b) Gin”* because the word “gin” makes up part of the word “ginnery”. The wording of the first part of the question, *“If beer is made in a brewery…”*, might also lead participants to conclude that the answer is likely to do with alcohol. However, it is possible that the wording of the question and the obviousness of *“(b) Gin”* might cause participants to elaborate on the question and the multiple-choice alternatives before giving their response, perhaps reasoning that the question is too easy and suspecting that there could be a trick. However, without any additional information about the correct answer, they may be unable to come up with a reason to select another alternative with any confidence. Therefore, they choose the obvious answer *“(b) Gin”* but assign it low confidence. However, unlike Outcome D in Figure 1, the participant has good understanding of the question and is highly attentive. Consequently, the process of elaboration before responding may result in corrective feedback becoming highly memorable despite the low confidence assigned to the original error.

In summary, we predicted that we would observe a hypercorrection effect for control questions, where high-confidence errors are more likely to be corrected than low-confidence errors. However, for trick questions, we predicted a non-monotonic relationship between confidence and error correction. In particular, if participants fail to see the trick, they will select the obvious answer and assign it high confidence and the surprise associated with those high-confidence errors will result in increased attention to and retention of the feedback. In contrast, trick questions that lead participants to be suspicious may lead to low-confidence, elaborated errors that would also show a high rate of error correction. Critically, though, although we believe similar mechanisms drive error correction in the latter case (i.e., enhanced attention to and retention of the feedback), it occurs without a metacognitive mismatch. Error correction for medium-confidence errors may be lower than either of these high- and low-confidence error-correction rates.

**Experiment 1**

Experiment 1 was our initial attempt at gathering evidence bearing on whether corrections to elaborated, low-confidence errors would be remembered well. In addition to comparing the error-correction rates for difficult control questions and trick questions as explained above, we attempted to experimentally manipulate confidence using a procedure reported by Koriat, Lichtenstein, and Fischhoff (1980). Specifically, for one set of questions, participants were required to justify their answer with reason(s) for why they selected it. For the other set, no justification was necessary. Koriat et al. found that generating reasons for answers augmented confidence because evidence supporting the answer was selectively retrieved in an attempt to justify it (i.e., confirmatory processing). For our purposes, this was a straightforward procedure to implement to experimentally manipulate confidence so that we could examine the effect of confidence on error correction without concern for item-selection problems.

 During the final testing phase, we also wanted to investigate what participants actually remembered about the corrective episode associated with making an error. Research has shown that children (Peeck & Tillema, 1978; 1979; Peeck, van den Bosch, & Kreupeling, 1985) and adults (Butler et al., 2011; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Vaughn & Rawson, 2012; Yan, Yu, Garcia, & Bjork, 2014) typically show good memory for initial test responses if they successfully correct errors. Furthermore, Fazio and Marsh (2009) found that source memory for feedback was enhanced when feedback to errors was surprising compared to unsurprising. Following on from this research, we asked our participants to recall their original responses and the corrective feedback from the initial test after responding to each final-test item.

**Method**

 **Participants.** Participants were 21 students (8 male) with a mean age of 23.29 years (*SD* = 3.66), from the University of Southampton who participated in return for course credit. One participant was excluded from all statistical analyses for failing to follow instructions, leaving 20 participants in total.

**Design.** The primary independent variable in this experiment was question type (control, trick). We also asked participants to provide a reason for half of their responses in order to experimentally manipulate confidence (Koriat et al., 1980).

Therefore, the experiment was a 2 x 2 within-subjects design with two independent variables, question type (control, trick) and whether a response required a reason (reason, no reason). The assignment of the reason vs. no-reason prompt to question-set halves in each condition (control, trick) was counterbalanced across participants. Although filler items were not analyzed, the same counterbalancing routine was used for assigning the reason prompt to them as well. The primary dependent variable was the proportion of errors on the initial test that were corrected on the final test. We also measured accuracy on both tests and confidence for correct and incorrect responses on the first test only. During the final test, we measured the recall proportion of both original errors and the feedback provided to errors.

**Materials and apparatus**. We used two types of general-knowledge questions, control and trick. All questions were presented in multiple-choice format with four alternatives. Firstly, a list of 20 moderately hard, control questions was constructed using an updated version of Nelson and Narens (1980) general-knowledge norms (Tauber, Dunlosky, Rawson, Rhodes, & Sitzman, 2013). Questions were selected if they produced a high proportion of errors accompanied by moderate confidence rating (i.e., not at floor) and we constructed the multiple-choice options ourselves. For example, “*What is the last name of the astronomer who published in 1543 his theory that the Earth revolves around the sun? (a) Copernicus (b) Galileo (c) Newton (d) Hubble*” [correct answer: “*(d) Copernicus”*].

We also included 10 trick questions, which were compiled using a variety of online resources. For example, “*How many of each type of animal did Moses take onto the Ark? (a) 0 (b) 1 (c) 2 (d) 4*” is an example of a trick question because the story of the Ark involves Noah, not Moses. Hence, the correct answer is “*(a) 0”* rather than the “obvious”, but wrong, answer *“(c) 2”.* Since this was our first experiment using trick questions, we erred on the side of caution by including only half as many trick questions as control questions. Bottoms, Eslick, and Marsh (2010) showed that participants were more likely to detect errors in general knowledge questions when such errors were more common compared to when they were rare. Therefore, in order to avoid participants getting suspicious about the nature of the questions or the experiment, which may have led them to respond strategically as the experiment went on (e.g., deliberately avoiding selection of the most obvious answer), we kept the number of trick questions lower than the number of control questions.

A further 10 easy filler questions were also included in the experiment, to avoid the task being so difficult that participants simply stopped trying to answer the questions, as well as four practice questions at the start. These were all taken from Tauber et al.’s (2013) updated general-knowledge norms and were chosen for their high proportion of correct responses. Our materials are shown in Appendix A.

The task was coded using LiveCode programming software and was administered on Apple iMac computers at individual workstations.

 **Procedure.** The experiment consisted ofa lab-based computer task. In the first phase, after completing four practice items, participants were presented with 40 general-knowledge questions, each accompanied by four multiple-choice options. They were instructed to choose one alternative and then rate their confidence that their answer was correct using a slider on a 25-100% scale, where 25% and 100% represented guessing and certainty, respectively. For half of the questions, participants were also asked to provide a reason for their answer, whereas for the other half, no reason was necessary. The computer randomized the presentation order of the questions and the order of the multiple-choice alternatives within each question for each participant.

Immediate corrective feedback was provided after completing each question. Participants were required to click a “Check Answer” button to view the feedback. They could not press the button until they had selected an answer, given a reason (if required), and provided a confidence rating. Once they chose to check their answer, they were not able to go back to change any of their responses and they were not able to move forward without viewing the feedback. The feedback consisted of verification (correct/incorrect) followed by the correct answer regardless of whether the participants’ response was correct or incorrect. For example, for the control question “*What is the last name of the astronomer who published in 1543 his theory that the Earth revolves around the sun? (a) Copernicus (b) Galileo (c) Newton (d) Hubble*” participants would be presented with *“Correct! The correct answer is…Copernicus”* or *“Incorrect! The correct answer is…Copernicus”* depending on the accuracy of their response. The whole procedure, including viewing of the feedback, was self-paced. Participants were required to click a button to remove the feedback message and move to the next question.

 When all 40 questions had been answered, participants completed an unrelated 15-to-20-minute facial recognition filler task. In the final phase of the experiment, participants were presented with the same 40 questions as in phase one, but this time in cued-recall format (e.g., “*What is the last name of the astronomer who published in 1543 his theory that the Earth revolves around the sun?* \_\_\_\_\_\_\_\_”). The order of the questions was randomized again for each participant. In addition, they were also asked whether they could recall their original response and the corrective feedback to each question. No confidence ratings were required in this phase. The whole task took approximately 45-60 minutes to complete. At the end of the session, participants were debriefed and thanked for their time.

**Results and Discussion**

Only data from control and trick questions were analyzed because, as expected, final test performance for the easy filler questions was at ceiling (*M* = .99, *SD* = .05). All dependent variables on both Test 1 and Test 2 were first analyzed with separate 2 (control, trick) x 2 (reason, no reason) repeated-measures Analyses of Variance (ANOVAs). The reason manipulation produced no main effects or interactions on any of the dependent variables, largest *F*(1, 19) = 3.20, *p* =.09, ηp2 = .10. This result did not coincide with our expectation that confidence would be significantly higher for items that required a reason compared to items that did not. Upon closer inspection, the failure to obtain an effect of the reason manipulation was almost certainly due to participants’ limited responding when asked to provide reasons for their responses. For example, a large proportion of participants simply wrote “I just guessed” or “I don’t know” for multiple items instead of providing a logical reason for their choice as per the instructions. Because the reason manipulation produced entirely null results, the data were pooled across this factor so that only one variable, question type, remained.

**Test 1 (multiple choice).** The number of errors on the first test was divided into seven bins according to confidence level (i.e., 25-40%, 41-50%, 51-60%, 61-70%, 71-80%, 81-90%, 91-100%). These data are shown in Table 1 as a function of question type. The upper panel in Table 2 shows mean accuracy, as well as mean confidence in both correct responses and errors, on Test 1 for both question types (control, trick). Separate one-way repeated-measures ANOVAs on these Test-1 dependent variables indicated that confidence for both correct and incorrect responses was significantly higher for trick questions compared to control questions, *F*(1,19) = 16.37, *p* = .001, ηp2 = .47, and, *F*(1,19) = 95.49, *p* < .001, ηp2= .83, respectively. In contrast, there was no difference in accuracy between the question types, *F* < 1.

**Test 2 (cued recall).** The bottom panel of Table 2 shows the means for the dependent variables used on Test 2 (accuracy, proportion of Test-1 errors corrected, proportion of Test-1 errors recalled, and proportion of Test-1 feedback to errors recalled).Again, each dependent variable was analyzed with separate one-way ANOVAs with question type as the factor.Whereas there was no difference in accuracy between control and trick questions on the first test, accuracy was significantly higher for trick questions on the final test, *F*(1,19) = 5.18, *p* = .035, ηp2 = .21. Consistent with this observation, participants corrected significantly more errors for trick questions than control questions, *F*(1,19) = 6.03, *p* = .024, ηp2 = .24. Furthermore, participants were significantly more likely to remember their original error, *F*(1,19) = 12.16, *p* < .01, ηp2 = .39, and the corrective feedback, *F*(1,18) = 40.20, *p* < .001, ηp2 = .69, for trick questions compared to control questions. These results are consistent with other research demonstrating that successful error correction is associated with good memory for original errors (e.g., Butler et al., 2011; Knight et al., 2012; Peeck & Tillema, 1978, 1979; Peeck et al., 1985; Vaughn & Rawson, 2012; Yan et al., 2014). Together, these results suggest that there is something about answering a trick question incorrectly and then being corrected that was particularly memorable.

We reasoned above that control questions would be likely to show a monotonic relationship between error correction and confidence whereas trick questions would show a non-monotonic relationship. To investigate this possibility, we divided participants’ responses on the 25-100% confidence scale into three bins: low (25-40), medium (41-70), and high (71-100).[[3]](#footnote-3) Figure 2 shows the proportion of errors corrected on the final test as a function of confidence and question type. A 3 (confidence: low, medium, high) x 2 (question type: control, trick) repeated-measures ANOVA on these data revealed that neither the main effect of question type nor confidence was significant, *F* < 1, and, *F*(2, 12) = 2.23, *p* = .15, ηp2 = .27, respectively. However, there was a significant interaction between the factors, *F*(2, 12) = 6.52, *p* =.01, ηp2 = .52. [[4]](#footnote-4) As expected for control questions, there was a monotonic relationship between error correction and confidence, which indicated a clear hypercorrection pattern. This pattern was confirmed statistically by calculating within-subjects gamma correlations between confidence and the proportion of errors corrected.[[5]](#footnote-5) A one-sample t-test showed that the mean gamma correlation (*M* = .20) was significantly different from zero, *t*(18) = 2.81, *p* = .012, *d* = .64.

In contrast, the trick questions produced a distinct V-shape in Figure 2, with low-confidence errors being just as likely to be corrected as high-confidence errors. These results suggest that the superior error-correction rate for trick questions compared to control questions was not only due to wrong answers to trick questions being assigned higher confidence as the hypercorrection effect would suggest. Instead, there was also a significant difference in error correction between the question types for *low*-confidence answers as well, such that the proportion of errors corrected for low-confidence trick questions was significantly higher than for low-confidence control questions, *F*(1, 11) = 8.86, *p* = .013, ηp2 = .45. The rates of error correction at the medium- and high-confidence levels did not differ significantly, both *F*s < 1.

**Summary.** Our results showed that accuracy was comparable across question types at Test 1, but at Test 2, accuracy was significantly higher for trick questions than control questions. This Test-2 accuracy difference was partially due to a significantly higher proportion of errors corrected on the final test for trick questions compared to control questions (Table 2).

 To explore the relationship between confidence and error correction in more detail, we plotted the error-correction rate as a function of confidence for each type of question. The results of this analysissupported our prediction that control questions would show the typical monotonic (hypercorrection) relationship between confidence and error correction whereas trick questions would not. Instead, we found that the likelihood of error correction for trick questions was equated for high- and low-confidence errors and both of these types of errors were more likely to be corrected on a final test than medium-confidence errors. Note that good error correction for the trick questions assigned low confidence occurred in the absence of a metacognitive mismatch, suggesting that feedback does not always have to be surprising for participants to attend to it and retain it.

Our findings suggest that there may be a qualitative difference in the types of errors committed between control and trick questions. In particular, we suspect that for some trick questions, participants were aware that they were being tricked but were unable to put their finger on what the trick actually was, even after some form of elaborative cognitive processing. Therefore, they chose the “obvious” (but wrong) answer because they were unable to deduce the true correct answer, assigning their response low confidence. We suspect that feedback for these types of questions was well remembered because the process of elaboration before selecting an answer increased attention to the feedback, which potentiated learning even though confidence was low and there was no metacognitive mismatch. In Experiment 2, we sought more direct evidence for these speculations.

**Experiment 2**

Experiment 2 was designed to further investigate error correction of low-confidence responses to trick questions. We reasoned that the trick questions we used in Experiment 1 were not necessarily homogenous because there are a number of ways in which a question could be considered “tricky”. For example, Moses-Illusion-type questions are those for which a critical piece of (mis)information within the question is often overlooked leading participants to give the obvious but incorrect answer (Erikson & Mattson, 1981). For example, the question “*What is the capital of Czechoslovakia? (a) Riga (b) Prague (c) Zagreb (d) None of the above*” is considered a Moses Illusion trick question because Czechoslovakia (which no longer exists as a country) is related to its true counterpart, Czech Republic. If this critical misinformation is missed, then participants are likely to confidently but incorrectly choose “*(b) Prague”*, which is the capital of the Czech Republic, instead of the correct answer “*(d) None of the above”*.

However, other trick questions often require participants to engage in some kind of lateral thinking in order to respond correctly. For example, the question “*Which cheese is made backwards? (a) Swiss (b) Gruyère (c) Cheddar (d) Edam”* does not contain any misinformation. In fact, all the information required to answer this question is presented within it, but participants must take a more indirect and creative approach if they are to respond correctly. In this example, they must think about the wording literally in order to arrive at the correct alternative (i.e., the word *made* is literally *Edam* spelled backwards).

In an attempt to gain some control over potential trick-question heterogeneity in Experiment 2, we conducted an online survey to gather norms data for some of the question set used in Experiment 1, plus some additional items (see Method section for details). In total, the survey included 42 questions, 28 of which (12 control, 10 trick, and 6 filler) were previously used in Experiment 1. In addition to these items, we included 14 new trick questions, which we gathered from a variety of online sources. One-hundred-and-fifty-six participants answered these 42 trick, control and filler multiple-choice questions. Critically, in addition to gathering a response to each item, we gave participants the opportunity to provide a second answer if they had low confidence in their first answer and asked them to provide their reasoning for both answers. This procedure allowed us to identify trick questions for which the trick was likely to have been missed (high-confidence, single-response items) and trick questions that may have been associated with some elaboration prior to answering the question (low-confidence, double-response items).

Importantly, if high elaboration prior to answering some of the double-response trick questions potentiates learning by increasing attention to the corrective feedback as we hypothesize, then trick questions of this type (selected a priori) should show excellent levels of error correction, despite the low confidence assigned to them and the absence of a metacognitive mismatch. Answers to some of the high-confidence, single-response trick questions may also show high levels of error correction but for a partially different reason: these questions also facilitate high attention to the feedback, but this occurs because of a metacognitive mismatch rather than high elaboration. Finally, some questions from both the double- and single-response question sets may be assigned medium confidence or low confidence because of confusion, lack of knowledge, and/or inattention (Outcome D in Figure 1). We do not expect good error correction for these items. By increasing the number of trick questions from Experiment 1 and including some low-confidence, double-response items, we were able to address the issue of trick-question heterogeneity as well as concerns surrounding the small numbers of observations at the low and medium confidence points in Experiment 1, thus allowing us to draw firmer conclusions.

The second main reason for conducting Experiment 2 was to gather more direct evidence that some low-confidence errors to trick questions are associated with high elaboration prior to answering the questions. To do so, we added three dependent variables to the design. First, we measured *response latency*, with the expectation that elaboration would cause the double-response trick questions to take longer to answer than either control questions or single-response trick questions. Second, participants were required to indicate how *surprised* they were and also whether they *understood* the corrective feedback. We expected that high elaboration would mean that double-response trick questions would show the least surprise, but the highest level of feedback understanding, compared to control or single-response trick questions. We made these predictions because we believed participants were expecting their low-confidence, double-response answers to trick questions to be wrong, but that elaboration prior to answering the question would make the feedback easily comprehensible.

**Method**

 **Participants.** Participants were 39 undergraduates (10 male) with a mean age of 19.67 years (*SD* = 3.25) from the University of Southampton. They were recruited through the university’s research website and took part in exchange for course credit.

 **Design.** This experiment was a repeated-measures design with one independent variable, question type (control, single-response trick, double-response trick). The primary dependent variable was the proportion of errors on the initial test that were corrected on the final test. In addition to the variables we measured in Experiment 1, we also measured response latency, level of surprise (5-point scale with 5 = high surprise), and whether the corrective feedback was understood (Y/N).

 **Materials and apparatus.** We used similar materials in this experiment as in Experiment 1 with some modifications. Firstly, we increased the number of trick questions from 10 (in Experiment 1) to 24. Of these, 16 trick questions were double-response questions. For these items, participants who took part in the online survey provided two responses and generally indicated low confidence. For example, consider the question “*If beer is produced in a brewery, what is produced in a ginnery? (a) Cotton (b) Gin (c) Wool (d) Vodka”.* The correct answer to this question is “*(a) Cotton”,* but the majority of participants answered “*(b) Gin*”, indicating low confidence and also providing a second response, perhaps because it seemed too obvious but they were unsure of the correct answer.

The remaining eight trick questions were single-response items for which participants of the online survey were seemingly unaware of any trick, indicating high confidence to the obvious (but incorrect) answer and not opting to provide a second response. For example, “*Which month has 28 days? (a) February (b) March (c) June (d) All of these”.* The majority of participants answered “*(a) February”* with high confidence because February is notorious for having only 28 days. However, the correct answer is “*(d) All of these”* because all months have (at least) 28 days. As well as trick questions, we also included 12 moderately hard control questions and eight easy filler questions, which were also used in Experiment 1. Four practice items were presented prior to the initial test as in Experiment 1. See Appendix B for the complete list of questions used in Experiment 2.

The task was coded using LiveCode programming software and was administered on Apple iMac computers at individual workstations.

 **Procedure.** The general procedure was the same as in Experiment 1 with the inclusion of some additional measures. After viewing the feedback to each question, participants were asked *“On a scale of 1 to 5, where 1 indicates ‘not at all surprised’ and 5 indicates ‘very surprised’, how surprised are you by the correct answer to this question?”*. They were required to select one number on the scale and were then asked *“Do you understand the feedback?”*, to which they responded either Yes or No. In addition, the program recorded the time it took participants to select an answer and rate their confidence for each question. Response latency was measured as the time from initial presentation of the question to the point at which the participant pressed the “Check Answer” button to indicate they had decided on their answer and rated their confidence. Participants were unaware that their response latencies were being measured. The final test consisted of the same questions that were presented in the initial test but in cued-recall format, as in Experiment 1. The whole experiment was self-paced, including viewing of the feedback during the initial test. The final alteration was that the filler task was changed to a 10-minute video of Milgram’s (1963) experiment on obedience followed by approximately five minutes of math problems.

**Results and Discussion**

Only data from control and trick questions were analyzed because, as expected, final test performance for the easy filler questions was at ceiling (*M* = .99, *SD* = .04).

**Test 1 (multiple choice).** The bottom panel of Table 1 shows the numbers of errors on Test 1 in each of seven confidence bins (i.e., 25-40%, 41-50%, 51-60%, 61-70%, 71-80%, 81-90%, 91-100%) as a function of question type. The top panel of Table 3 shows the means for each of the dependent variables used in Test 1 (accuracy, confidence for correct answers, confidence for errors, response latency for errors, surprise rating for errors, and proportion of feedback to errors understood). We conducted separate one-way, repeated-measures ANOVAs on each of the Test-1 dependent variables and found that question type exerted significant effects on all of them [accuracy: *F*(2, 76) = 12.20, *p* < .001, ηp2 = .24; confidence for correct answers, *F*(2, 68) = 35.41, *p* < .001, ηp2 = .51; confidence for errors, *F*(2, 76) = 128.83, *p* < .001, ηp2 = .77; surprise, *F*(2, 76) = 22.17, *p* < .001,ηp2= .37; response latency, *F*(2, 76)= 11.65, *p* < .001, ηp2 = .24; and feedback understanding, *F*(2, 76) = 22.36, *p* < .001, ηp2 = .37]. As Table 3 shows, accuracy was highest for the control questions. However, confidence for both errors and correct responses was highest for the single-response trick questions and lowest for the double-response trick questions.

As expected, the response latency to answer double-response trick questions was the highest of all the question types, which suggests that participants elaborated on these questions before answering them.[[6]](#footnote-6) This elaboration for double-response trick questions seemingly helped to make the corrective feedback understandable as well, particularly when compared to single-response trick questions, which showed the lowest level of feedback understanding. The finding that surprise ratings were highest for single-response trick questions supports this conclusion. Elaboration of double-response trick items may have also contributed to lower surprise ratings for these questions compared to the single-response trick questions because participants would have been more expectant of incorrect feedback. Together, the results from Test 1 suggest that our pre-experimental survey was successful at identifying the two types of trick questions.

**Test 2 (cued recall).** We turn now to the Test-2 data with particular interest in how error correction on Test 2 is related to confidence on Test 1. The bottom panel of Table 3 shows the means for each of the dependent variables used in Test 2 (accuracy, proportion of Test-1 errors corrected, proportion of Test-1 errors recalled, and proportion of Test-1 feedback to errors recalled). Although Test-1 accuracy was highest for control questions, and lowest for single-response questions, there were no differences in accuracy on Test 2, *F*(2, 76) = 1.53, *p* = .22, ηp2 = .04. Equal Test-2 accuracy derived primarily from the high error-correction rate of the double-response trick questions; despite their having the lowest level of Test-1 confidence, they produced the highest level of error correction, *F*(2, 76) = 3.88, *p* = .025, ηp2 = .09, which resulted, descriptively at least, in the highest level of Test-2 accuracy.

Participants’ Test-1 errors to trick questions were also remembered better than errors to control questions, *F*(2,76) = 34.66, *p* < .001, ηp2=.48, for the main effect. This was particularly true of high-confidence responses, but even the low-confidence errors were recalled reasonably well. Also, the feedback for double-response trick questions was remembered best, and much better than feedback for control questions, *F*(2,76) = 16.03, *p* < .001, ηp2 = .30, for the main effect. Clearly, this pattern of results is inconsistent with any proposal that high confidence in initial errors and a metacognitive mismatch is required for good memory for details about Test-1 and high levels of error correction.

As in Experiment 1, we next split participants’ confidence ratings into three bins: low (25-40), medium (41-70), and high (71-100) to further examine the relationship between confidence and error correction within the question sets. However, in order to statistically analyze the data in this manner, and to rectify the issue of low numbers of observations at the low- and medium-confidence points that occurred in Experiment 1, it was necessary to pool together the single- and double-response trick questions conditions. Figure 3 shows the proportion of errors corrected on the final test as a function of confidence across question types. The pattern of error correction for control questions followed the same pattern as in Experiment 1; as confidence increased, error correction increased monotonically, although in this experiment, the mean gamma correlation (*M* = .13) was not significantly different from zero, *t*(19) = 1.25, *p* = .23, *d* = .28. More critically, the relationship between confidence and error correction for trick questions was clearly very different from the control questions. In particular, Figure 3 shows that for trick questions, the *highest* level of error correction corresponded to the *lowest* level of confidence.

 A 3 (confidence: low, medium, high) x 2 (question type: control, trick) repeated-measures ANOVA on these data revealed that neither the main effect of question type nor confidence was significant, *F*(1, 24) = 2.47, *p* = .13, ηp2 = .09, and *F* < 1, respectively.[[7]](#footnote-7) These results were similar to Experiment 1, but unlike Experiment 1, the interaction was not significant, *F*(2, 48) = 1.07, *p* = .35, ηp2 = .04. However, the critical data correspond to those errors assigned the lowest confidence: we assume it is specifically low-confidence errors to trick questions that are associated with high elaboration and good error correction. Consistent with this assumption, errors to trick questions at the lowest level of confidence were more likely to be corrected than lowest-confidence control questions, *F*(1, 33) = 5.17, *p* = .030, ηp2 = .14. The proportion of errors corrected at the medium- and high- confidence levels were comparable across question types, *F* < 1 in both cases.

**Multiple regressions.** The fact that we measured a number of additional variables in this experiment, including level of surprise, response latency, and level of understanding of the feedback, allowed us to investigate error correction in greater depth. Specifically, the proportion of errors corrected for control, single-, and double-response trick questions were entered as the dependent variables in three, separate stepwise multiple regressions. Confidence for errors, surprise rating, response latency, proportion of feedback understood, proportion of Test-1 errors recalled and the proportion of feedback to errors recalled were entered as the independent variables to determine, which, if any, of these variables were significant predictor(s) of error correction. If high levels of error correction were driven by confidence and surprise due to metacognitive mismatches as the hypercorrection effect suggests, then we would expect that confidence and surprise ratings would emerge as the strongest predictors of error correction, at least for control questions. The results are shown in Table 4.

As Table 4 shows, a significant model emerged for control questions, *F*(1, 38) = 49.23, *p* < .001, *R*2 = .57. There was only one significant predictor of error correction: whether the participant could recall the feedback to their errors from Test 1. A significant model for single-response trick questions was also observed, *F*(2, 38) = 7.29, *p* = .002, *R*2 = .29. For these questions, there were two significant predictors of error correction: whether the participant could remember the feedback to errors from the initial test and whether they could remember their original errors. Finally, we found a significant model for the double-response trick questions, *F*(1, 38) = 30.09, *p* < .001, *R*2 = .45. Similar to control questions, the only significant predictor of error correction was whether the participant could remember the feedback to errors from the initial test. Thus, across all three regressions, recall of feedback and recall of Test-1 errors were the only significant predictors of error correction. Critically, neither confidence nor level of surprise, factors thought to underpin metacognitive mismatches and therefore important to error correction, accounted for any unique variance in any model, not even the one for control questions.

 **Summary.** The results ofExperiment 2 provide additional supporting evidence that there is more to error correction than metacognitive mismatches, at least with some types of material. Of all the question types, the double-response trick questions, which were selected a priori based on our survey results, ranked the *lowest* in terms of both accuracy and confidence on the initial test. Despite this poor performance that was reflected in confidence, these questions were associated with the *highest* rate of error correction and consequently ranked the highest in terms of accuracy on the final test.

Examining error correction for each question type as a function of confidence also largely replicated the pattern observed in Experiment 1 for control questions, although the distinct V-shape for trick questions observed in Experiment 1 (Figure 2) was not replicated for trick questions in Experiment 2 (Figure 3). In particular, error correction for control questions was monotonically related to confidence, such that the rate of error correction increased with confidence. For trick questions there was also a monotonic relationship between confidence and error correction; however, it was in the opposite direction. Thus, for trick questions, error correction was highest at the lowest confidence level and lowest and the highest confidence level.

The results of the multiple regression analyses showed that neither confidence nor surprise was a significant predictor of error correction. Instead, the proportion of errors corrected was mostly predicted, for both types of questions, by whether participants were able to recall the feedback they were given in the initial testing phase. For single-response trick questions, the proportion of errors recalled was also a significant predictor; however, the relationship between these two variables was negative. This latter result was unexpected because we thought that being tricked into making a high-confidence error would be a particularly memorable event, leading to good memory for both the original error and the corrective feedback (e.g. Butler et al., 2011; Knight et al., 2012; Peeck & Tillema, 1978; 1979; Peeck et al., 1985; Vaughn & Rawson, 2012; Yan et al., 2014). However, it appears from our findings that error correction for these items, at least in part, relies on poor memory for the original error.

This finding seems at odds with Butterfield and Metcalfe (2001) who reported that participants reproduced their original errors on a final test when asked to give three possible answers, but were able to discriminate between the error and the correct response. However, the difference between our results and theirs could be due to either the different materials used in our study and the nature of trick questions and/or the use of multiple-choice questions on Test 1 that led to high levels of interference (i.e., participants were better off forgetting their original error so that it didn’t interfere with memory for the corrective feedback). Since we are yet to find a definitive explanation for the negative relationship between memory for original errors and error correction, we feel it may warrant further investigation. For present purposes, the most important take-home message from the multiple regression analyses is that both confidence and surprise failed to predict error correction in any of our analyses. This result again supports the notion that metacognitive mismatches are not necessary to produce good error correction.

**General Discussion**

In two experiments, we investigated error correction and its relationship to confidence. In both experiments, error correction for hard general-knowledge (control) questions was monotonically related to confidence. That is, high-confidence errors were more likely to be corrected than medium-confidence errors, which, in turn, were more likely to be corrected than low-confidence errors – the typical finding observed in the hypercorrection literature. However, error correction for trick questions showed a very different relationship to confidence. In particular, low-confidence errors for trick questions were just as likely to be corrected on the final test as high-confidence errors. In Experiment 1, the relationship between confidence and error correction was a non-monotonic V-shape, with both high- and low-confidence errors being more likely to be corrected than medium-confidence errors. However, with the inclusion of an increased number of trick questions in Experiment 2, the relationship was monotonic, but *reversed* compared to control questions (i.e., low > medium > high; see Figure 3).

We also found other evidence demonstrating excellent error correction for low-confidence errors. For example, double-response trick questions used in Experiment 2 were associated with the highest level of error correction, despite having the lowest level of confidence on the first test. The high rate of error correction for these questions meant that their accuracy went from lowest of the three types of questions on the first test to the highest on the final test (Table 3). Finally, regression analyses in Experiment 2 indicated that neither confidence nor surprise was a significant predictor of error correction, a result that would be expected if both low- and high-confidence errors were retained well (albeit for different reasons). Instead, error correction was predicted by recall of feedback for all question types, and for the single-response trick questions, previous errors as well.

We hypothesized that the good error correction for low-confidence errors to trick questions was attributable to considerable elaboration (problem solving) occurring prior to answering the question, which potentiated learning of the feedback. High elaboration prior to committing low-confidence errors was supported in Experiment 2 by response latencies as well as by the feedback-understanding and surprise ratings. In particular, compared to single-response trick questions, for which the trick was likely missed, participants took longer to produce errors to low-confidence trick questions, but indicated that they were less surprised at being corrected and understood the feedback more. This is exactly the pattern of results one might expect if participants were undecided about which one of two or more options was correct but had spent some time considering evidence for and against each alternative. Overall, these results indicate that, under some circumstances and with the right materials, high confidence is not necessary for corrective feedback to be encoded and remembered well, and that there are certain types of material for which the hypercorrection pattern (high-confidence error correction > low-confidence error correction) does not generalize.

This excellent correction of low-confidence errors is not consistent with the traditional understanding in the literature that such errors simply reflect misunderstanding, inattention, low knowledge, or confusion (i.e., Outcome D in Figure 1). Undoubtedly, sometimes low-confidence errors conform to this description, which would lead to a poor error correction rate on subsequent tests. Indeed, it seems likely that there were some low-confidence errors of this type even with the trick questions used in our research. However, our research also shows that, under the right circumstances, low-confidence errors may not just be random guesses, but may actually reflect a reasonably high level of understanding due to elaboration. This elaboration and understanding is helpful in retaining feedback and correcting errors.

The possibility that there may be more than one type of low-confidence error suggests there may need to be a fifth pathway (ending in a fifth outcome) added to Figure 1 to capture error correction with trick questions. Similar to random low-confidence errors, that pathway would begin with low-confidence incorrect responses made on Test 1. However, if the material is familiar to participants and elaboration occurs prior to answering the questions, then good rather than poor Test-2 performance may be observed (Outcome E). In the next section, we consider some potential mechanisms for such an outcome.

**Mechanisms for Good Correction of Highly-Elaborated, Low-Confidence Errors**

We have argued that good error correction for highly-elaborated, low-confidence errors might be caused by an increase in attention to the feedback. That is, after a period of elaboration, participants become interested in knowing the correct answer to the question and pay close attention to the feedback once it is presented, enhancing its retention in memory. Such an account is similar to other accounts of good error correction that occur in cases of metacognitive mismatches (i.e., Outcome B and C in Figure 1; Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009). For example, Kulhavy (1977; Kulhavy et al., 1976) suggested that when feedback is discrepant with expectations, participants spend longer processing the feedback, which aids its retention in memory. If enhanced attention also applies to highly-elaborated, low-confidence errors, it suggests that attention to feedback can be heightened for reasons other than there being a violation of expectations (i.e., a metacognitive mismatch).

 However, there may be at least one other possible explanation for good error correction of highly-elaborated, low-confidence errors that we have not yet considered. Elaboration prior to answering the question may activate a number of knowledge structures and memories relevant to the question. Activation of these structures and memories may, in turn, facilitate encoding of the feedback once it is presented. By this account, it is not the attention to the feedback that is the primary cause of good error correction but facilitated integration of the feedback into current knowledge.

This account is not dissimilar from the “knew-it-all-along” explanation of the hypercorrection effect (e.g., Butterfield & Mangels, 2003; Metcalfe & Finn, 2011, Sitzman et al., 2014, 2015), which suggests that prior knowledge is predictive of error correction rather than confidence per se, such that when participants have more prior knowledge of the topic of a question, feedback processing is enhanced. However, out explanation may also differ in important ways from the “knew it all along” account. First, the elaboration we have described is not necessarily linked to prior domain knowledge per se. Instead, it is likely caused by characteristics of particular questions such as closely competing answer alternatives or metaknowledge that there may be a trick and that the most obvious answer is probably wrong. These factors likely persuade participants to elaborate further before answering, thereby activating memories pertinent to that specific question, which enhances feedback encoding. Importantly, this question-specific elaboration may not be dependent on high amounts of prior knowledge as the “knew it along” account suggests.

A second difference between the accounts is that, in most cases, if participants have high prior domain knowledge, they also respond with high confidence. This almost perfect confounding of high prior domain knowledge and high confidence has made it a challenge to determine which of these two factors is the cause of the hypercorrection effect. In contrast, as we have noted throughout, selections of the most obvious answer for highly-elaborated trick questions in our research were mostly made with low, not high, confidence. Thus, our experiments may be some of the first to unconfound confidence from the beneficial effects on feedback encoding caused by activation of memories related to the question. In this way, trick questions, and perhaps other types of questions that lead people to query whether there is “more than meets the eye” after a superficial reading, may be a way forward to disentangling the effects of confidence versus prior memory on error correction.

**Retrieval Effort**

 Our results are consistent with other research that has highlighted the importance of retrieval effort (i.e., elaboration) in promoting memory. For example, Pyc and Rawson (2009) found that an elaborative, difficult retrieval attempt was more likely to promote retention of the correct information compared to facile retrieval.[[8]](#footnote-8) Similarly, Jacoby (1978) had participants study related cue-target word pairs by either reading intact pairs or constructing the target themselves using the letter prompts that were provided. He found that reaching a solution by constructing it led to better retention of the target than simply reading it. He argued that reading a given solution is an effortless event that does not serve to enhance retention, whereas the conscious effort associated with solving a problem facilitates memory for the target.

Kornell, Klein, and Rawson (2015) proposed a two-stage framework of elaborative retrieval in which Stage 1 is the retrieval attempt and Stage 2 is processing of the target. According to this model, which was designed primarily to explain the effect of retrieval and feedback on memory for cue-target word pairs, a process of elaboration or searching for the answer occurs at Stage 1, during which semantic associates of the cue word are activated. Stage 1 is completed, and Stage 2 begins, when the target word becomes available, either through successful retrieval or an external source (e.g., feedback). In Stage 2, both direct and indirect links between the cue and target are strengthened, whereas irrelevant links are weakened. Kornell et al. argued that retrieval success is not critical to retention of the target and error correction on later tests. Rather, it is the process of elaboration during Stage 1 of retrieval that facilitates memory of the correct target.

Richland et al. (2009) also showed that failed retrieval attempts were beneficial to learning. They found evidence of a *pretesting effect* such that performance on a final test was significantly higher when participants incorrectly answered questions about to-be-studied material before actually studying it. Furthermore, Arnold and McDermott (2013) found that tests, even when unsuccessful, potentiated subsequent learning by improving the efficacy of later study. Richland et al. argued that failed retrieval attempts facilitate deep processing of the question (e.g. Bjork, 1975), where participants actively engage in an elaborative search for the correct answer. Furthermore, they suggested that such elaborated errors “may create a fertile ground for later encoding of the answer” (p. 252).

Similar to Richland et al. (2009), Arnold and McDermott (2013), and Kornell et al. (2015), we argue that, actually, it is not necessary for retrieval attempts to be successful because elaborative cognitive processing prior to making a low-confidence error potentiates learning of the corrective feedback. This result builds on the findings that elaborative encoding before the first *correct* retrieval attempt leads to improved retention for the correct response (Karpicke & Smith, 2012) and that elaborative questioning facilitates learning (e.g., Pressley, McDaniel, Turnure, Wood, & Ahmad, 1987). In line with these findings and Bjork’s (1994) principle of desirable difficulty, we propose that the more effort that is involved in producing an error that is later corrected, to trick questions at least, the more likely it is to stay corrected (for a review of the principle of desirable difficulty, see Bjork & Bjork, 2011).

**Pure Guessing and Error Correction**

Some recent research on error correction has shown that corrective feedback to very low-confidence errors (i.e., pure guesses) can be retained. For example, Kornell et al. (2009) found that guessing the answers to fictional general-knowledge questions (Experiments 1 and 2) and guessing weak associates to single words (Experiments 3-6) followed by corrective feedback led to better performance on a subsequent test than reading the questions and answers together. Although they did not measure confidence, it is likely that confidence was low to the generated answers. In perhaps an even more striking example, Potts and Shanks (2014; see also Yang, Potts, & Shanks, 2017) showed participants either very rare English words or words from a foreign vocabulary, asked them to guess the meanings, and then provided corrective feedback. In Potts and Shanks’ own words, the novelty of their materials would have meant that their participants “…would have known that their generated guesses were almost certain to be wrong” (p. 660). However, compared to a condition in which the intact pairs were read, generated errors were corrected on a subsequent test leading to higher accuracy for those items, despite the low confidence.

How do we reconcile the finding of good error correction of complete guesses on the one hand with the hypercorrection literature demonstrating the importance of high confidence to error correction on the other? One issue to consider is the control condition that is used in these different literatures. In research on the hypercorrection effect, the comparison is typically between high- versus low-confidence errors. In contrast, in the literature on guessing correction, the control condition is usually a read- or study-only condition. Thus, even though confidence ratings are not commonly gathered in the guess-correction literature, one could envisage that error correction has the following pattern under most circumstances: high-confidence errors > low-confidence errors > read only. However, although this pattern captures the typical scenario, our research has suggested that sometimes, with the right materials, the pattern can be low-confidence errors ≥ high-confidence errors > read only.

Interestingly, in addition to the read and generate groups, Potts and Shanks also included a “choice” group who were provided with the correct answer and a lure and instructed to guess the correct answer. Across four experiments, final test performance on these items was no greater than intact items that were simply read. At first blush, these choice-group results would appear to be at odds with our findings demonstrating excellent error correction for low-confidence, multiple-choice trick questions. However, in our view, the discrepancy can easily be explained by considering how familiar participants were with the materials. As noted above, Potts and Shanks’ materials were completely novel to participants whereas ours were not. Consequently, it seems unlikely that participants in Potts and Shanks’ study would have, or even could have, elaborated on the question prior to answering it. In other words, the errors that participants made in their choice conditions likely corresponded to low-elaboration, low-confidence errors (Outcome D in Figure 1) rather than highly-elaborated, low-confidence errors as seen with our trick questions (Outcome E as per the discussion above). Without the high elaboration prior to making the error, the kind of potentiated feedback learning we observed with our trick questions would have been absent and, without it, processing of the corrective feedback would not have been enhanced.

**Conclusion**

Overall, our findings suggest that there are qualitative differences in the types of errors that are committed and that the hypercorrection effect does not generalize to all types of material. We replicated the hypercorrection effect using moderately hard multiple-choice questions but found that low-confidence errors to trick questions were just as likely to be corrected as high-confidence errors in both experiments. We argue that confidence (and surprise) may not be as critical in error correction as the existing hypercorrection literature suggests. Remembering corrective feedback is not just a matter of confidence but perhaps more intuitively, the ability to recall the feedback is what leads to successful error correction. Furthermore, we suggest that elaborative cognitive processing of the question and multiple-choice alternatives before selecting a response could be a key to successful retention of corrective feedback to low-confidence errors, such that it promotes increased attention to feedback similar to the metacognitive mismatch frequently reported in the hypercorrection literature. Our findings highlight that elaboration is a powerful tool for remembering corrective feedback and this finding certainly warrants further investigation. We hope that our findings provide a good foundation upon which a new model of error correction could potentially emerge.

References

Anderson, J. R, & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*, 186-197. doi:10.1037/0096-3445.128.2.186

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940-945. doi:10.1037/a0029199

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.) *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York: Worth Publishers.

Bjork. R.A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.) *Information processing and cognition: The Lojtola Symposium* (pp. 123-144). Hillsdale. NJ: Erlbaum.

Bjork, R.A. (1994). Metamemory and memory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.) *Metacognition: Knowing about knowing* (pp. 85-205). Cambridge, MA: MIT press.

Bottoms, H. C., Eslick, A. N., & Marsh, E. J. (2010). Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences. *Memory*, *18,* 670-678. doi:10.1080/09658211.2010.501558

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. doi:10.3758/s13423-011-0173-y

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

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

Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research, 17*, 793-817. doi:10.1016/s0926-6410(03)00203-9

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

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

Elley, W. B. (1966). The role of errors in learning with feedback. *British Journal of Educational Psychology*, *36*, 296-300. doi:10.1111/j.2044-8279.1966.tb01882.x

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*, 540-551. Doi:10.1016/S0022-5371(81)90165-1

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

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

Guthrie, J. T. (1971). Feedback and sentence learning. *Journal of Verbal Learning and Verbal Behavior, 10*, 23-28. doi:10.1016/S0022-5371(71)80088-9

Higham, P.A. (2015, November). *New improved gamma: Using the area under the ROC curve to indirectly compute a more accurate Goodman-Kruskal gamma coefficient.* Paper delivered at the 54th Annual Meeting of the Psychonomic Society*,* Chicago, IL.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649-667. doi:10.1016/S0022-5371(78)90393-6

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

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*, 17-29. doi: 10.1016/j.jml.2012.02.004

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*, 254-284. doi:10.1037/0033-2909.119.2.254

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*, 731-746. doi:10.1016/j.jml.2011.12.008

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107-118. doi:10.1037/0278-7393.6.2.107

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. doi:10.1037/a0015729

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 283. doi:10.1037/a0037850

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*, 211-232. doi:10.3102/00346543047002211

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 68*, 505-512. doi:10.1037/h0033243

Kulhavy, R. W., & Parsons, J. A. (1972). Learning-criterion error perseveration in text materials. *Journal of Educational Psychology, 63*, 81-86. doi:10.1037/h0032348

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279-308. doi:10.3102/00346543047002211

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

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79-97. doi:10.3102/00346543058001079

Melton, A. W., & Irwin, J. McQ. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology, 53,* 173-203. doi:10.2307/1417415

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: did they know it all along? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 437-448. doi:10.1037/a0021962

Muller, G. E., & Schumann, F. (1894). Experimentelle beiträge zur untersuchung des gedächtmisses. *Zeitschrift Psychology*, *6*, 81-90.

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition, 3*, 222-229. doi:10.1016/j.jarmac.2014.05.001

Nicol, D. J., & Macfarlane‐Dick, D. (2006). Formative assessment and self‐regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199-218. doi: 10.1080/03075070600572090

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56*, 132-143. doi:10.1037/h0057488

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8. doi:10.1037/0278-7393.31.1.3

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187-193. doi:10.3758/BF03194050

Peeck, J., & Tillema, H. H. (1978). Delay of feedback and retention of correct and incorrect responses. *The Journal of Experimental Education*, *47*, 171-178. doi:10.1080/00220973.1978.11011677

Peeck, J., & Tillema, H. H. (1979). Learning from feedback: Comparison of two feedback procedures in a classroom setting. *Perceptual and motor skills*, *48*, 351-354. doi: 10.2466/pms.1979.48.2.351

Peeck, J., van den Bosch, A. B., & Kreupeling, W. J. (1985). Effects of informative feedback in relation to retention of initial responses. *Contemporary Educational Psychology, 10,* 303–313. doi: 10.1016/0361-476X(85)90028-1

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143,* 644–667. doi:10.1037/a0033194

Pressley, M., McDaniel, M. A., Turnure, J. E., Wood, E., & Ahmad, M. (1987). Generation and precision of elaboration: Effects on intentional and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 291–300. doi:10.1037/0278-7393.13.2.291

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437-447. doi:10.1016/j.jml.2009.01.004

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. doi:10.1037/a0016496

Sitzman, D. M., Rhodes, M. G., & Tauber, S. K. (2014). Prior knowledge is more predictive of error correction than subjective confidence. *Memory and Cognition, 42*, 84-96. doi:10.3758/s13421-013-0344-3

Sitzman, D. M., Rhodes, M. G., Tauber, S. K., & Liceralde, V. R. T. (2015). The role of prior knowledge in error correction for younger and older adults. *Aging, Neuropsychology, and Cognition*, *22*, 502-516. doi: 10.1080/13825585.2014.993302

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General-knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods*, *45*, 1115-1143. doi:10.3758/s13428-012-0307-9

Webb, L. W. (1917). Transfer of training and retroaction: A comparative study. *Psychological Monographs, 24*, 1-90. doi:10.1037/h0093121

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*, 899-905. doi:10.3758/s13423-012-0276-0

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition*, *42*, 1373-1383. doi:10.3758/s13421-014-0454-6

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance Online Publication. doi:10.1037/xlm0000363

|  |
| --- |
| Table 1 |
| *The Numbers of Errors for the Different Question Types on the Initial Test (Test 1) as a Function of Confidence Rating in Experiments 1 and 2* |
| Experiment and Question Type | Confidence | Total |
| 25-40 | 41-50 | 51-60 | 61-70 | 71-80 | 81-90 | 91-100 |
| Experiment 1 |
|  | Control | 91 | 48 | 13 | 36 | 26 | 25 | 20 | 259 |
|  | Trick | 24 | 9 | 7 | 7 | 6 | 12 | 61 | 126 |
|  | Total | 115 | 57 | 20 | 43 | 32 | 37 | 81 | 385 |
| Experiment 2 |
|  | Control | 94 | 36 | 15 | 24 | 28 | 19 | 34 | 250 |
|  |  |  |  |  |  |  |  |  |  |
|  | Single-response Trick | 14 | 7 | 12 | 15 | 14 | 30 | 146 | 238 |
|  |  |  |  |  |  |  |  |  |  |
|  | Double-response Trick | 185 | 78 | 34 | 37 | 59 | 29 | 52 | 474 |
|  |  |  |  |  |  |  |  |  |  |
|  | Total | 293 | 121 | 61 | 76 | 101 | 78 | 232 | 962 |

Table 2

*Mean Results [±95% Confidence Intervals] for All Dependent Variables As a Function of Question Type and Test in Experiment 1.*

|  |  |
| --- | --- |
|  | Question Type |
|  Dependent Variable and Test | Control | Trick |
| Test 1 (Multiple Choice) |  |
| Accuracy | .35[.29, .42] | .37[.29, .45] |
| Confidence for correct answers (%) | 69.26[62.92, 76.81] | 85.28[76.81, 93.75] |
| Confidence for errors (%) | 53.78[48.53, 59.04] | 76.44[71.52, 81.35] |
| Test 2 (Cued Recall) |  |
| Accuracy  | .62[.52, .72] | .72[.62, .82] |
| Proportion of errors corrected | .51[.41, .62] | .66[.54, .78] |
| Proportion of Test-1 errors recalled  | .77[.69, .84] | .88[.82, .93] |
| Proportion of Test-1 feedback to errors recalled  | .64[.53, .76] | .91[.86, .95] |

*Note*. For all dependent variables except accuracy at initial test, the difference between conditions was statistically significant to at least *p* < .05.

Table 3

*Mean Results [±95% Confidence Intervals] for All Dependent Variables As a Function of Question Type and Test in Experiment 2*

|  |  |
| --- | --- |
|  | Question Type |
|  | Control |  | Trick |
| Test and Dependent Variable | All |  | Single Response | Double Response |
|  Test 1 (Multiple Choice) |  |  |  |
| Accuracy | .36[.32, .40] |  | .25[.19, .31] | .24[.21, .27] |
| Confidence for correct answers (%) | 61.57[55.62, 67.51] |  | 87.49[81.76, 93.22] | 59.29[53.71, 64.87] |
| Confidence for errors (%) | 55.87[51.78, 59.97] |  | 86.93[84.23, 89.63] | 54.03[49.90, 58.15] |
| Response latency for errors (sec) | 12.79[11.62, 13.96] |  | 13.44[11.85, 15.04] | 15.70[14.31, 17.08] |
| Surprise rating for errors (/5) | 2.90[2.67, 3.12] |  | 3.77[3.57, 3.97] | 3.20[2.95, 3.45] |
| Proportion of feedback to errors understood (Y/N) | .88[.81, .95] |  | .66[.58, .75] | .81[.74, .89] |
| Test 2 (Cued Recall) |  |  |  |
| Accuracy | .67[.61, .73] |  | .68[.61, .75] | .72[.67, .78] |
| Proportion of Test-1 errors corrected | .60[.52, .68] |  | .60[.52, .67] | .69[.63, .75] |
| Proportion of Test-1 errors recalled  | .42[.38, .47] |  | .69[.62, .76] | .52[.47, .57] |
| Proportion of Test-1 feedback to errors recalled  | .39[.34, .44] |  | .52[.46, .58] | .55[.52, .59] |

Table 4

*Unstandardized and Standardized Regression Coefficients for Significant Predictors of Error Correction in Each Model for Control and Trick Questions*

|  |  |
| --- | --- |
|  | Regression Co-efficient |
| Variable | B | SE b | Β |
| Control |  |  |  |
| Proportion of feedback to Test-1 errors recalled  | 1.15 | .16 | .76 \*\*\* |
| Single-Response-Trick |  |  |  |
| Proportion of feedback to Test-1 errors recalled  | 1.24 | .32 | .98 \*\* |
| Proportion of Test-1 errors recalled | -.84 | .28 | -.79 \*\* |
| Double-Response-Trick |  |  |  |
| Proportion of feedback to Test-1 errors recalled  | .98 | .18 | .67 \*\*\* |
|  |

FigureCaptions

*Figure 1.* Tree diagram depicting the impact that corrective feedback received on a first test (Test 1) has on subsequent (Test-2) performance as a function of the accuracy and confidence associated with the initial Test-1 response. Feedback has little impact on Outcome A. However, memory for the feedback is facilitated in cases where there is a metacognitive mismatch (Outcomes B and C), which benefits Test-2 performance. Finally, feedback has little impact on Outcome D if the low-confidence associated with this outcome is due to misunderstanding, low knowledge, and/or inattention.

*Figure 2.* Mean proportion of errors corrected as a function of confidence and question type in Experiment 1. Values at each point on the graph represent the number of observations at each confidence level. Error bars represent *SEM*s.

*Figure 3.* Mean proportion of errors corrected as a function of confidence and question type in Experiment 2. Values at each point on the graph represent the number of observations at each confidence level. Error bars represent *SEMs*.

Appendix A

Experiment 1 Question Set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Question Type | Question | Correct Answer | Alternative 1 | Alternative 2 | Alternative 3 |
| Control | What is the last name of the man who began the reformation in Germany? | Luther | Hitler | Mendelssohn | Habsburg |
| What is the capital of Canada? | Ottawa | Toronto | Edmonton | Quebec |
| What is the name of the largest desert on Earth? | Antarctica | Sahara | Arabian | Kalahari |
| What is the name of the unit of measure that refers to a six-foot depth of water? | Fathom | League | Knot | Cable length |
| What is the capital of New York? | Albany | New York | Buffalo | Rochester |
| What is the last name of the astronomer who published in 1543 his theory that the Earth revolves around the sun? | Copernicus | Galileo | Newton | Hubble |
| What is the name for a cyclone that occurs over land? | Tornado | Hurricane | Monsoon | Sirocco |
| What is the last name of the most popular pin-up girl of World War II? | Grable | Monroe | Hayworth | Lake |
| What is the longest river in South America? | Amazon | Mississippi | Colorado | Danube |
| What is the capital of Australia? | Canberra | Sydney | Perth | Melbourne |
| What is the name of the brightest star in the sky excluding the sun? | Sirius | North | Pollux | Vega |
| In what European city is the Parthenon located? | Athens | Rome | Paris | None of the above |
| Which planet was the last to be discovered? | Neptune | Pluto | Uranus | None of the above |
| What is the name of the mountain range that separates Asia from Europe? | Ural | Himalayas | Andes | Sayan |
| What is the proper name for a badminton bird? | Shuttlecock | Birdie | Ball | Feather |
| What is the last name of the man who invented the telegraph? | Morse | Bell | Edison | Watson |
| For which country is the Yen the monetary unit? | Japan | China | Thailand | Bangladesh |
| Of which country is Baghdad the capital? | Iraq | Afghanistan | Saudi Arabia | Syria |
| What is the last name of the criminal who was known as Scarface? | Capone | Montana | Manson | Lombardo |
| What is the last name of the first flier to fly solo around the world? | Post | Earheart | Gatty | Taylor |
| Trick | Which country invaded Ethiopia in 1935? | None of the above | Italy | France | Great Britain |
| How many of each type of animal did Moses take onto the Ark? | 0 | 1 | 2 | 3 |
| Which month has 28 days? | All of the above | February | March | June |
| What is the capital of Czechoslovakia? | None of the above | Prague | Riga | Zagreb |
| What was the highest mountain in the world before Mount Everest was discovered? | Everest | Makalu | Lhotse | K2 |
| If you were running a race and passed the person in second place what place are you in now?  | Second | First | Third | All of the above |
| Which cheese is made backwards? | Edam | Swiss | Cheddar | Gruyere |
| Divide 50 by half and add 20. | 120 | 45 | 35 | 140 |
| In which direction is the smoke going if an electric train is travelling south?  | None of the above | North | South | East |
| The national flag of France consists of three horizontal stripes of red blue and which other colour?  | None of the above | White | Green | Yellow |
| Filler | What is the capital of the UK? | London | Cardiff | Manchester | Edinburgh |
|  | What is the name of Tarzan's girlfriend? | Jane | Belle | Jasmin | Ariel |
|  | What is the name of the rubber object that is hit back and forth by ice hockey players? | Puck | Ball | Disc | Jack |
|  | What is the name given to a giant ocean wave cause by an earthquake? | Tsunami | Swell | Undercurrent | Billow |
|  | What name is given to a doctor who specialises in cutting open the body? | Surgeon | Paediatrician | Dermatologist | None of the above |
|  | What is the name for the long sleep some animals go through during the entire winter? | Hibernation | Slumber | Hypersomnia | None of the above |
|  | Which supposedly unsinkable ship sunk on its maiden voyage in 1912? | Titanic | Arosa Star | Fort Victoria | Oceanos |
|  | Which precious gem is red? | Ruby | Emerald | Topaz | All of the above |
|  | Who wrote Romeo and Juliet?  | Shakespeare | Chaucer | Tennyson | Williams |
|  | What is the name of the horse-like animal with black and white stripes? | Zebra | Donkey | Hippopotamus | Llama |
| Practice | Who is the current Prime Minister of the UK? | David Cameron | Nick Clegg | Ed Milliband | Nigel Farage |
|  | Which country is the world's biggest producer of coffee? | Brazil | USA | South Africa | Argentina |
|  | The term 'pulmonary' relates to which organ? | Lungs | Kidneys | Liver | Brain |
|  | In Rudyard Kipling’s 'The Jungle Book' what kind of creature is Kaa? | Snake | Bear | Human | Tiger |

Appendix B

Experiment 2 Question Set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Question Type | Question | Correct Answer | Alternative 1 | Alternative 2 | Alternative 3 |
| Control | What is the capital of Canada?  | Ottawa | Toronto | Quebec | Edmonton |
| What is the name for a cyclone that occurs over land?  | Tornado | Hurricane | Monsoon | Sirocco |
| What is the longest river in South America?  | Amazon | Mississippi | Danube | Colorado |
| In what European city is the Parthenon located?  | Athens | Rome | Paris | None of these |
| For which country is the Yen the monetary unit?  | Japan | China | Bangladesh | Thailand |
| Of which country is Baghdad the capital?  | Iraq | Afghanistan | Saudi Arabia | Syria |
| What is the name of the largest desert on Earth?  | Antarctica | Sahara | Arabian | Kalahari |
| What is the capital of New York?  | Albany | New York | Buffalo | Rochester |
| What is the last name of the first flier to fly solo around the world?  | Post | Earheart | Gatty | Taylor |
| What is the last name of the man who invented the telegraph?  | Morse | Edison | Bell | Watson |
| Single-Response Trick | The national flag of France consists of 3 horizontal stripes of red blue and which other colour? | None of these | White | Green  | Yellow |
| Divide 50 by half and then add 20. | 120 | 45 | 35 | 140 |
| If an electric train is travelling south in which direction is the smoke blowing? | None of these | North | South | West |
| How many of each type of animal did Moses take onto the Ark? | 0 | 1 | 2 | 4 |
| Which month has 28 days? | All of these | February | March | June |
| Who coined the phrase "survival of the fittest"? | Herbert Spencer | Charles Darwin | John Fleming | Charles Doppler |
| If you were running a race and passed the person in second place what place are you in now? | Second | First | Third | None of these |
| Double-Response Trick | If beer is made in a brewery what is produced in a ginnery?  | Cotton | Gin | Wool | Vodka |
|  | What is the longest animal in the world? | Ribbon worm | Blue whale | Reticulated python | Giant oarfish |
|  | What colour was the sky in Ancient Greece? | Bronze | Blue | White | Gold |
|  | What is the capital of Czechoslovakia? | None of these | Prague | Zagreb | Riga |
|  | Which of these cities is south of Florence - Italy? | Marseille - France | Bologna - Italy | Vienna - Austria | Zagreb - Croatia |
|  | Which of these is correct - "penguins flies" or "A penguin flies"? | Neither | A penguin flies | Penguins flies | Both |
|  | Where would you find Robert McCulloch's London Bridge? | USA | UK | France | Mexico |
|  | Autophobia is the fear of what? | Being alone | Cars | Illness | Wax statues |
|  | What was the Prime Minister's name in 1980? | David Cameron | Margaret Thatcher | Harold Wilson | Tony Blair |
|  | How much dirt is there in a hole that measures 2ft x 3ft x 4ft? | 0 cubic ft | 60 cubic ft | 6 cubic ft | 12 cubic ft |
|  | Which of these cities is north of Toronto - Canada? | Venice - Italy | New York - USA | Detroit - USA | Madrid - Spain |
|  | Which country invaded Ethiopia in 1935? | None of these | Italy | France  | Great Britain |
|  | What was the highest mountain in the world before Everest was discovered? | Everest | K2 | Lhotse | Makalu |
|  | What does a muscologist study? | Mosses | Muscles | Sugar | Grapes |
|  | Which cheese is made backwards? | Edam | Swiss | Cheddar | Gruyere |
|  | In Ancient Greece what animal would you have expected to see in a hippodrome? | Horse | Hippo | Cow  | Rhino |
| Easy Filler | What is the capital of the UK? | London | Cardiff | Manchester | Edinburgh |
|  | What is the name of Tarzan's girlfriend? | Jane | Belle | Jasmin | Ariel |
|  | What is the name of the rubber object that is hit back and forth by ice hockey players? | Puck | Ball | Disc | Jack |
|  | What name is given to a doctor who specialises in cutting open the body? | Surgeon | Paediatrician | Dermatologist | None of the above |
|  | What is the name for the long sleep some animals go through during the entire winter? | Hibernation | Slumber | Hypersomnia | None of the above |
|  | Which precious gem is red? | Ruby | Emerald | Topaz | All of the above |
|  | Who wrote Romeo and Juliet?  | Shakespeare | Chaucer | Tennyson | Williams |
|  | What is the name of the horse-like animal with black and white stripes? | Zebra | Donkey | Hippopotamus | Llama |
|  | What is the capital of the UK? | London | Cardiff | Manchester | Edinburgh |
| Practice | Which country is the world's biggest producer of coffee? | Brazil | USA | South Africa | Argentina |
|  | The term 'pulmonary' relates to which organ? | Lungs | Kidneys | Liver | Brain |
|  | In Rudyard Kipling’s 'The Jungle Book' what kind of creature is Kaa? | Snake | Bear | Human | Tiger |
|  | Which tree bears conkers? | Horse Chestnut | Oak | Silver Birch | Willow |

1. The outcomes in Figure 1 do not necessarily apply if the delay between tests is long (e.g., Butler, Fazio & Marsh, 2011). [↑](#footnote-ref-1)
2. More recently, some research has shown that sometimes feedback provided to pure guesses can aid learning compared to reading (e.g., Kornell, Hays & Bjork, 2009; Potts & Shanks, 2014). We consider this research more fully in the General Discussion. [↑](#footnote-ref-2)
3. We reduced the number of bins from seven (shown in Table 1) to three to avoid losing participants who did not contribute data to all cells in the design. For the same reason, we also used only three bins rather than seven for the analogous analysis in Experiment 2. [↑](#footnote-ref-3)
4. The means presented in Figure 2 are based on all of the available data for each cell. However, it was necessary to drop several participants from the ANOVA pertaining to Figure 2 because of missing data (e.g., not all participants contributed data to all levels of confidence for both question types). Consequently, the results from the ANOVA should be interpreted with some caution. The number of participants contributing data to the ANOVA and other analyses can be determined by the degrees of freedom presented with each analysis. [↑](#footnote-ref-4)
5. Higham (2015) showed that *Ag* provides a more accurate estimate of association than the more common Goodman-Kruskal gamma co-efficient if the latter is computed with the usual formula using discordant and concordant pairs of observations. Consequently, we first computed *Ag* and then transformed it to gamma using the formula 2*Ag* – 1. [↑](#footnote-ref-5)
6. Note that the difference in response latency is not due to participants producing two responses as they did in the online survey. In Experiment 2, participants always provided only one response to all questions. [↑](#footnote-ref-6)
7. The means presented in Figure 3 are based on all of the data available for each cell. However, as in Experiment 1, it was necessary to drop participants from the ANOVA who did not contribute data to all cells. The number of participants contributing data to the ANOVA and other analyses can be determined from the reported degrees of freedom. [↑](#footnote-ref-7)
8. One difference between our research and that of Pyc and Rawson (2009) is that our focus was on the effect of elaboration prior to making errors, whereas their research focused on elaboration prior to successful retrieval. [↑](#footnote-ref-8)