MEMORY

ISSN: 0965-8211 (Print) 1464-0686 (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

Receiving right/wrong feedback: Consequences for learning

Lisa K. Fazio , Barbie J. Huelser , Aaron Johnson & Elizabeth J. Marsh

To cite this article: Lisa K. Fazio , Barbie J. Huelser , Aaron Johnson & Elizabeth J. Marsh (2010) Receiving right/wrong feedback: Consequences for learning, MEMORY, 18:3, 335-350, DOI: 10.1080/09658211003652491

To link to this article: https://doi.org/10.1080/09658211003652491

MEMORY, 2010, 18 (3), 335–350

Psychology Press
Taylor & Francis Group

Receiving right/wrong feedback: Consequences for learning

Lisa K. Fazio, Barbie J. Huelser, Aaron Johnson, and Elizabeth J. Marsh
Duke University, Durham, NC, USA

Prior work suggests that receiving feedback that one's response was correct or incorrect (right/wrong feedback) does not help learners, as compared to not receiving any feedback at all (Pashler, Cepeda, Wixted, & Rohrer, 2005). In three experiments we examined the generality of this conclusion. Right/wrong feedback did not aid error correction, regardless of whether participants learned facts embedded in prose (Experiment 1) or translations of foreign vocabulary (Experiment 2). While right/wrong feedback did not improve the overall retention of correct answers (Experiments 1 and 2), it facilitated retention of low-confidence correct answers (Experiment 3). Reviewing the original materials was very useful to learners, but this benefit was similar after receiving either right/wrong feedback or no feedback (Experiments 1 and 2). Overall, right/wrong feedback conveys some information to the learner, but is not nearly as useful as being told the correct answer or having the chance to review the to-be-learned materials.

Keywords: Feedback; Learning.

In many everyday situations learning is not errorless. The student who studies for an exam makes mistakes on the test, the partygoer mixes up the names of the other guests, and the American abroad becomes confused about the few words of French she learned for her trip. Given that a mistake has occurred, what steps should be taken to ensure the correct answer is retrieved in the future? One obvious suggestion is to provide *feedback* regarding the mistake.

Recently, Pashler and colleagues made an important contribution to this literature with a thorough examination of the memorial consequences of different types of feedback, over different delays (Pashler, Cepeda, Wixted, & Rohrer, 2005). Their participants learned Luganda–English word pairs (Luganda is a language spoken in Uganda). Immediately after studying the word pairs, participants took a cued recall test that required them to translate each of a series of Luganda words into English. After each translation was typed, the correct answer was presented to one third of participants (answer feedback), whereas others only learned whether a translation was correct or incorrect (right/wrong feedback). A third group received no information about the correctness of their answers (no feedback). A second cued recall test (also with feedback) immediately followed the first test. When a Luganda word had been translated correctly on the first test, the feedback manipulation had no consequences for performance on the second test. Of greater interest, however, was whether feedback helped participants to correct errors made on the first test. In this case answer feedback was key, leading to the highest performance on the

Address correspondence to: Lisa K. Fazio, Psychology & Neuroscience, Duke University, 9 Flowers Drive, Box 90086, Durham, NC 27708-0086, USA. E-mail: lisa.fazio@duke.edu

This work was supported by a collaborative activity award from the James S. McDonnell Foundation. We thank Mark McDaniel and Andrew Butler for useful comments on this research, and Katya Fernandez, Yusha Liu, Amy Hsu, Tyson Wepprich, and Lauren Rosenberg for preparation of stimuli, data collection, and scoring. The second author is now at Columbia University.

http://www.psypress.com/memory DOI:10.1080/09658211003652491

second test. Right/wrong feedback was unhelpful: participants who received right/wrong feedback performed no better than the participants who received no feedback. These results persisted to a final test a week later.

Pashler et al. (2005) clearly demonstrated that answer feedback was advantageous in learning situations such as the one they used, and that right/wrong feedback provided no more benefit than receiving no feedback at all. Our question, however, is whether right/wrong feedback can be beneficial in other situations. This is an important question to answer, because of the practical implications for teachers: giving correct answer feedback takes more time than does simply marking an answer as correct or incorrect. As described below, our goal was to identify situations where right/wrong feedback yielded benefits similar to those observed following answer feedback. Obtaining the same learning outcome with less instructor effort would be beneficial for both students and teachers.

In contrast to receiving no feedback, right/wrong feedback provides information to the learner, telling him/her whether a response is correct or incorrect. In this paper we explore three different ways in which receiving this information might benefit learning. The first is whether right/wrong feedback is beneficial with types of to-be-learned material other than foreign vocabulary. The second is whether right/wrong feedback promotes the retention of correct answers, since it provides information about correctness that no feedback does not. The third is whether right/wrong feedback can promote error correction, if students are later given the chance to find the correct answer. We describe these three possibilities below.

First, the usefulness of right/wrong feedback might depend on the nature of the to-be-learned material. Pashler et al. (2005) used very simple stimuli that were essentially paired associates: foreign vocabulary words paired with their English translations. However, Roper (1977) found right/wrong feedback to be more effective than no feedback when participants were working with basic statistics. In Roper's study, receiving right/wrong feedback was still less effective than answer feedback, but it clearly provided a benefit over no feedback. Thus it is important to examine right/wrong feedback paired with different types of materials, as the informational value of right/wrong feedback may be higher when students are working with more complex materials. Therefore, in Experiment 1 we used prose materials.

Second, regardless of change in the overall score on the final test, right/wrong feedback may selectively promote the retention of correct answers, since it provides information about correctness that no feedback does not. The literature is surprisingly silent on this question. One issue is that right/wrong feedback is often compared to answer feedback, rather than to no feedback (e.g., Swindell & Walls, 1993; Waldrop, Justen, & Adams, 1986). Another issue (also noted by Pashler et al., 2005) is that the data are often not conditionalised on responses on the initial test, a prerequisite for examining benefits to correct answers. Guthrie (1971) found no added benefit from feedback following correct answers, but he examined only answer feedback and not right/wrong feedback. It is possible that the results would be different with right/wrong feedback. The participants receiving answer feedback have more important information to direct their attention to, namely using the feedback to correct erroneous responses. Thus, the participants receiving right/wrong feedback may direct more attention to the "correct" feedback simply because they are not focused on learning new information. The only relevant finding with right/wrong feedback (that we know of) is Pashler et al.'s experiment with foreign vocabulary, where right/wrong feedback had no benefit for correct answers. Therefore additional research is needed. In Experiment 1 we examined whether right/wrong feedback improves retention of facts correctly recalled from prose passages. In Experiments 2 and 3 we examined retention of correct translations of foreign vocabulary, with a focus on *low-confidence* correct translations. Given that low-confidence correct responses benefit from being paired with answer feedback (Butler, Karpicke, & Roediger, 2008), right/wrong feedback might serve a similar function.

We turn now to the third possible benefit of right/wrong feedback: that while right/wrong feedback may be ineffective on its own, it may be effective when paired with later review of the to-be-learned material. The real-world parallel is actually a common one: post-exam review is standard educational practice. For example, a student might re-read her textbooks or lecture notes after receiving her graded exam, to understand her grade and to prepare for later tests in the class. The first question of interest is whether the student benefits more from reviewing these

materials if she previously received right/wrong feedback than if she received no feedback at all. Having received right/wrong feedback should guide the student's review so that she pays more attention to things that she got wrong on the exam, which should benefit later memory. In other words, compared to a no feedback condition, students should learn more from review when armed with the information about what they do and do not know.

A second question is how review paired with right/wrong feedback will compare to having received answer feedback. Careful review effectively means the student receives answer feedback, except that the learner discovers the correct information when rereading the material instead of simply being told the correct answer. What is interesting about this situation is that it should allow students the benefits of answer feedback (exposure to the correct answer) while making the learning process less passive. The literature on *desirable difficulties* predicts that students may benefit from this more active search for feedback than from simply reading the feedback. Specifically, the idea of desirable difficulties is that challenging a learner can enhance later performance (Bjork, 1994; Schmidt & Bjork, 1992). For example, long-term retention is generally better after testing than re-studying (e.g., Glover, 1989; Roediger & Karpicke, 2006), after answering short-answer questions rather than multiple-choice questions (Kang, McDermott, & Roediger, 2007), and following spaced rather than massed practice (e.g., Dempster, 1988; Melton, 1970). In all of these cases students expend extra effort, which leads to better retention of the material. The ''desirable'' part of the label is important: not all difficulties are desirable ones. Students must succeed at challenges during learning in order for the challenges to be desirable. The question asked here is whether having to find the answers on one's own (while reviewing the original to-be-learned material) is a desirable difficulty, with positive memorial consequences. Forcing students to find the correct answers for themselves, following right/wrong feedback, may lead to better performance than simply telling the students the correct answers.

We could only find one paper that crossed feedback form (answer, right/wrong, or none) with review (allowed or not). However, in that experiment, learning in the review condition was to criterion and involved re-testing, making it hard to interpret the results for our purposes (Wentling, 1973). Thus it is an open question whether right/wrong feedback provides information to the learner that can guide later review, with positive consequences for memory.

In short, we examined the value of right/wrong feedback, to see if right/wrong feedback (a) facilitated the learning of material from prose passages, (b) helped correct answers to persist (and whether this depended on confidence in the correct answer), and (c) aided the correction of errors given a chance to review. Prose materials were examined in Experiment 1, and the question of correct answer persistence was examined in all three experiments. Correct answer persistence was also specifically linked to confidence in initially correct answers in Experiments 2 and 3. The question of whether right/wrong feedback paired with review would facilitate error correction was examined in Experiments 1 and 2.

EXPERIMENT 1

In Experiment 1 we examined the interaction of feedback and review using prose materials. Participants read a series of non-fiction passages on topics such as ''Alaska'' and ''The Sun''. After reading, the participants took a cued recall test on the material. For some questions they received no feedback, for others right/wrong feedback, or answer feedback. Afterwards, half of the participants were allowed to review the passages for a second time. All participants then took a final test on the passages. In the no review condition we were interested in whether we replicated Pashler et al.'s (2005) results of the ineffectiveness of right/wrong feedback (as compared to the no feedback condition) and the ineffectiveness of feedback following correct answers. Within the review condition, of interest was whether participants would allocate more time to reviewing the sentences related to questions that they answered erroneously on Test 1, especially in the right/wrong feedback condition, and whether any extra time would translate into better performance on the final test. Longer review of sentences related to errors (as compared to correct answers or filler sentences) would provide evidence that participants used their review time strategically. Of particular interest was whether having received right/wrong feedback would promote strategic reviewing, as compared to having seen answer feedback or no feedback. Finally, we also manipulated whether or not items were tested

initially, to see if any benefits of reviewing depended on having been tested previously.

Method

Participants. A total of 48 Duke University undergraduates received monetary compensation for participating in the experiment. Participants were tested individually or in small groups of up to four people.

Design. The design was a 2 (initial test: tested, not tested) $\times$ 3 (feedback: none, right/wrong, answer) $\times$ 2 (review: none, review) mixed design. Initial testing and feedback condition were manipulated within-participants and review was a between-participants variable. The main dependent measure was the change in performance from Test 1 to Test 2. We also examined the mean time spent reviewing the passages in the review condition.

Materials. We took 12 non-fiction passages on history, geography, and science from Roediger and Marsh (2005); each passage contained four critical facts. The passages ranged in length from 11 to 24 sentences ($M = 14.33$). The passages were modified slightly so that a sentence never referred to more than one critical fact. Thus, in each passage, two sentences contained facts that were tested on the initial test and two sentences contained facts that were not tested initially. On average, almost 70% of sentences were filler sentences that did not contain critical facts. Both the passages and the questions are available from the corresponding author on request.

Across participants we counterbalanced which facts appeared on the first test; each fact was tested initially for half of the participants and not tested for the remaining participants. Additionally, the 12 passages were divided into three sets of four passages, to allow the counterbalancing of feedback condition. In other words, each participants received no feedback on responses to questions pertaining to four of the passages, right/wrong feedback for four other passages, and answer feedback for the remaining four passages. Across participants, all passages appeared equally often in the three feedback conditions.

In summary, there were 8 questions per feedback type on the initial test (summing to 24 questions, as the remaining 24 facts were not tested initially). The final test consisted of the questions on all 48 critical facts. The initial and final tests were both in cued recall format.

Procedure. The experiment involved five main parts: a study period, an initial test, a review period, a filled delay, and a final test.

First, during the *study period* participants read 12 different passages on the computer; the passages appeared in a set random order. Only one passage sentence appeared at a time, and participants pressed the spacebar to proceed to the next sentence. On average it took the participants about 17 minutes to read all 12 passages.

After reading all of the passages, participants took the *initial test*, which consisted of 24 questions in cued recall format. The passages were tested in the same order as they had been read (e.g., if the passage titled ''The Sun'' was studied first, questions on ''The Sun'' appeared first). Participants were encouraged to check their answers for misspellings to avoid errors in scoring and they were also warned against guessing.

Feedback was manipulated within participants, meaning each participant saw all three types of feedback on the initial test. For no feedback trials, participants received no feedback and simply proceeded to the next question after typing each response. For right/wrong feedback trials, the word ''correct'' or ''incorrect'' was displayed on the screen for 5 seconds before the computer advanced to the next question. ''Correct'' was displayed when the participant typed the correct answer or very close to the correct answer;[1] all other answers led to ''incorrect'' being displayed. On answer feedback trials, the correct answer was displayed for 5 seconds before the computer advanced to the next question, regardless of the participant's answer. Before the initial test, participants read instructions and completed three example questions that demonstrated each of the three feedback conditions.

Following the initial test was the *review phase.* Participants in the review condition reviewed all of the passages, sentence-by-sentence at their own pace. They did not have access to the initial test or their responses on that test; rather, they simply went through the original passages for a second time. Review was done on the computer and participants pressed a key whenever they were ready to move on to the next sentence. Participants in the no review condition worked on a paper-and-pencil visuo-spatial puzzle packet for 10 minutes instead of reviewing the passages.

[1] Common alternate misspellings were generated and the experiment was created so that the program would display ''correct'' for these responses.

The review period was immediately followed by a 20-minute *filled delay*, where participants either continued to work on the paper-and-pencil visuo-spatial puzzle packet (no review) or began to work on the puzzles (review).

Lastly, the *final test* consisted of the 48 critical questions in cued recall format. No feedback was provided. Participants were once again warned against guessing.

Results

All results are significant at the .05 level unless otherwise noted.

Change in proportion correct from the initial to the final test. Which conditions yielded the most improvement from the initial to the final test? For ease of presentation we answered this question using change scores as the dependent variable: for each participant we subtracted the proportion correct on the initial test from the proportion correct on the final test. Thus a positive change score reflects improvement across tests, and the question is whether improvement was equal in all conditions or whether it depended on feedback form and review. For completeness, the top portion of Table 1 shows performance on both the initial and final tests as well as the change scores used in the analyses. An examination of Table 1 reveals there was some variability in initial test performance (even though the feedback and review manipulations had not yet occurred), but the conclusions were the same when the analyses were computed on proportion correct (and time of test was used as a within-participants factor).

Replicating Pashler et al. (2005), answer feedback was the most helpful type of feedback. Items paired with answer feedback showed a larger improvement across tests ($M = .30$) than those paired with right/wrong feedback ($M = .09$) or no feedback ($M = .05$), $F(2, 92) = 20.64$, $MSE = .04$. Improvement in the right/wrong feedback and no feedback conditions was equally low, $t < 1$.

Reviewing the material was also effective, $F(1, 46) = 10.42$, $MSE = .07$. Participants who reviewed the material improved more across tests ($M = .22$) than those who were unable to review ($M = .08$). Finally, and most importantly, there was a significant interaction between type of feedback and review, $F(2, 92) = 4.38$, $MSE = .04$. Without review, only answer feedback improved performance. With review, there was improvement across tests in all three conditions. Review was most helpful for facts tested without feedback or paired with right/wrong feedback, as opposed to facts tested with answer feedback. Compared to

TABLE 1
Proportion correct on initial and final tests as a function of type of feedback and review condition

		No feedback	*Right/wrong feedback*	*Answer feedback*
Experiment 1				
No review	Initial test	.53	.48	.50
	Final test	.46	.47	.79
	Change	−.06	−.01	.30
Review	Initial test	.55	.54	.54
	Final test	.72	.73	.84
	Change	.17	.19	.30
Experiment 2				
No review	Initial test	.33	.37	.47
	Final test	.32	.39	.66
	Change	−.01	.02	.18
Review	Initial test	.42	.34	.39
	Final test	.80	.75	.87
	Change	.38	.41	.49
Experiment 3				
No review	Initial test	.28	.26	.30
	Final test	.25	.23	.42
	Change	−.03	−.03	.12

The materials were facts from prose passages in Experiment 1 and Luganda–English word pairs in Experiments 2 and 3. The standard error of the change score was .02 for Experiment 1, .03 for Experiment 2, and .01 for Experiment 3.

TABLE 2
Proportion correct on final test conditional on initial test answer, as a function of type of feedback and review condition

			No feedback	*Right/wrong feedback*	*Answer feedback*
Experiment 1	Answer on test 1:				
	Correct	No review	.77	.71	.85
		Review	.87	.88	.90
	Error	No review	.18	.31	.79
		Review	.68	.63	.85
Experiment 2	Answer on test 1:				
	Correct	No review	.87	.97	.88
		Review	.96	.95	.95
	Error	No review	.09	.12	.56
		Review	.76	.78	.83
Experiment 3	Answer on test 1:				
	Correct	No review	.77	.81	.79
	Error	No review	.03	.03	.28

The materials were facts from prose passages in Experiment 1 and Luganda–English word pairs in Experiments 2 and 3. For correct answers, the standard error was .03 for Experiment 1, .02 for Experiment 2, and .02 for Experiment 3. For errors, the standard error was .04 for Experiment 1, .05 for Experiment 2, and .02 for Experiment 3.

the no review condition, review boosted performance for facts tested without feedback, $t(46) = 3.57$, $SE = .06$, and facts paired with right/wrong feedback, $t(46) = 3.86$, $SE = .05$. For facts tested with answer feedback, review did not change improvement across tests, $t < 1$. In short, review led to improvement on facts paired with right/wrong or no feedback, although answer feedback still led to the highest performance.

Effects of feedback and review on initially correct answers. In order to more closely examine the effects of feedback and review we did a series of conditional analyses. We first examined items that were initially answered *correctly*. To preview, similar to Pashler et al. (2005), we found little effect of our manipulations on items initially answered correctly. The relevant data are shown in the top portion of Table 2. First, consider the effects of feedback condition (collapsing over whether or not review occurred). Items paired with answer feedback ($M = .88$) appeared to be answered correctly more often than those paired with right/wrong feedback ($M = .80$) or no feedback ($M = .82$), but this difference was not significant, $F(2, 92) = 2.09$, $MSE = .04$, $p = .13$. Next, consider the effects of prior review (collapsing over the different feedback conditions). There was a trend for reviewing to increase final performance ($M = .89$) compared to not reviewing the passages ($M = .78$), but this failed to reach traditional levels of significance, $F(1, 46) = 3.29$, $MSE = .13$, $p = .08$. There was no interaction between feedback and review, $F(2, 92) = 1.11$, $MSE = .13$, $p = .33$. Overall, most questions that were answered correctly on the initial test were also answered correctly on the final test, regardless of the type of feedback or the ability to review. There was no indication that right/wrong feedback promoted the retention of correct answers any more than receiving no feedback.

Effects of feedback and review on initial errors. We next examined the effects of feedback and review for questions that were initially answered erroneously. To be clear, ''erroneously'' means that participants produced a wrong answer, as opposed to answering ''I don't know''.[2] This analysis included only the 37 participants who provided data for all three cells.

When an error had been made on the first test, both review and feedback condition mattered. Regardless of review condition, errors were more likely to be corrected on the final test following answer feedback ($M = .82$) than right/wrong ($M = .47$) or no feedback ($M = .43$), $F(2, 70) = 14.46$, $MSE = .12$. Furthermore, regardless of feedback condition, participants who reviewed the material were more likely to correct their errors ($M = .72$) than were participants who were unable to review the material ($M = .43$), $F(1, 35) = 22.53$, $MSE = .11$. There was also an interaction between type of feedback and review, $F(2, 70) = 4.09$, $MSE = .12$. Although reviewing always led to better

[2] The data were also analysed conditional upon a ''don't know'' response, with results similar to those obtained when errors were analysed. We included only ''incorrect'' responses in this analysis to parallel the analysis in Pashler et al. (2005).

performance on the final test, this difference was magnified for items that had been paired with right/wrong or no feedback.

Review time. How did participants in the review condition spend their time during the review period? The top portion of Table 3 shows the average time (in seconds) spent reviewing each sentence, as a function of feedback condition and whether or not participants correctly produced the sentence's fact on the initial test. A 2 (answer on test 1: correct, error) × 3 (feedback condition: none, right/wrong, answer) within-participants ANOVA was computed. Of the 24 participants in the Review condition, 19 had observations in all six cells and were included in the analysis.

Overall, participants spent similar amounts of time reviewing the critical sentences, regardless of whether or not they had been able to correctly produce the fact on the initial test, $F < 1$. They spent an average of 3.2 seconds rereading each sentence containing a fact produced correctly on the initial test, similar to the 3.4 seconds spent on each sentence for which they had made an error.

Prior feedback condition also had no effect on time spent reviewing the sentences, $F(2, 36) = 1.31$, $p = .28$. Participants spent about 3.5 seconds reviewing sentences containing facts for which they had received no feedback or right/wrong feedback, and this did not differ from the time spent reviewing sentences containing facts paired with answer feedback. Finally, there was no interaction between prior feedback condition and correctness on the initial test, $F(2, 36) = 1.06$, $p = .36$. In short, participants' review times were largely insensitive to the manipulations of interest.

Testing effects. All of the analyses thus far have examined performance on questions that were tested on both the initial and final tests. The experimental design also included items that were tested for the first time on the final test. The manipulation of initial testing was included so that we could see if reviewing had benefits that went beyond the tested items.

First, was performance better on the final test for facts tested initially? More critically, did tested items benefit more from review than nontested items? To answer these questions, a 2 (initial test: tested, not tested) × 2 (review: none, review) mixed ANOVA was computed. As shown in Table 4, we found the standard testing effect: facts tested on the initial test were more likely to be answered correctly on the second test, $F(1, 46) = 10.78$, $MSE = .02$. Furthermore, participants who reviewed the passages performed better on the final test ($M = .70$) than those who did not review ($M = .56$), $F(1, 46) = 15.97$, $MSE = .03$. There was also a significant interaction between testing and review, $F(1, 46) = 4.10$, $MSE = .02$. The testing effect was not significant in the no review condition, even though it was present numerically (performance increased from 54% correct for not tested items to 57% following testing, $t(23) = 1.07$, $SEM = .03$, $p = .30$). The testing effect was only significant for participants who reviewed the material; prior testing boosted these participants' performance from the baseline of 63% to 77% correct, $t(23) = 3.28$, $SEM = .04$. Even so, review still benefited non-tested items, with participants in the review condition correctly

TABLE 3
Mean time (s) spent reviewing each sentence (Exp. 1) or word pair (Exp. 2) for items answered correctly or erroneously on the initial test, as a function of type of feedback received

	No feedback	*Right/wrong feedback*	*Answer feedback*	*M*
Experiment 1				
Answer on test 1:				
Correct	3.05	3.71	2.84	3.20
Error	3.93	3.46	2.92	3.44
Mean	3.49	3.59	2.88	
Experiment 2				
Answer on test 1:				
Correct	3.12	2.95	3.29	3.12
Error	8.05	9.17	8.58	8.60
Mean	5.56	6.06	5.94	

For correct answers, the *SE* was .31 for Experiment 1 and .28 for Experiment 2. For errors, the standard error was .44 for Experiment 1 and .81 for Experiment 2.

TABLE 4
Performance on the final test as a function of prior testing and review

	Not tested	*Tested*	*M*
No review	.54	.57	.56
Review	.63	.77	.70
Mean	.59	.67	

Data are from Experiment 1 (with prose materials). The standard error was .02 in the no review condition and .03 in the review condition.

answering 63% of final questions corresponding to non-tested items, as compared to 54% in the no review condition, $t(46) = 2.43$, $SE = .04$.

Did prior testing change review behaviour? We re-visited the review times (Table 3), focusing on time spent reviewing for tested vs non-tested facts. Participants spent marginally more time reviewing tested facts ($M = 3.04$ s) than non-tested facts ($M = 2.73$ s), $t(23) = 1.85$, $SE = .17$, $p = .08$. The reader will recall that review times were unaffected by ability to correctly produce the fact on the initial test, or the prior feedback condition. Participants were only sensitive to which items had been tested initially, and devoted more time to reviewing them. This differential attention to previously tested facts likely lead to the robust testing effect observed in the review condition (.77 vs .63) as compared to the non-significant one observed in the no review condition (.57 vs .54).

Discussion

The no review condition provided a nice conceptual replication of Pashler et al. (2005), using prose materials instead of foreign vocabulary. When review was not allowed, only answer feedback led to improved performance on the final test. Switching to prose materials did not change the conclusions about right/wrong feedback versus no feedback. Unlike Roper (1977), who used statistics materials, we found that right/wrong feedback was no more beneficial than receiving no feedback. Using prose materials also did not change the conclusions about the effects of feedback on correct answers: initially correct answers persisted to the final test at high rates and this was no greater following right/wrong feedback.

Reviewing prose passages helped students to learn the critical facts. However, review was no more helpful for items that had been paired with right/wrong feedback than those receiving no feedback. Even though right/wrong feedback informed participants about what they needed to look for in the review period, there was no evidence that they acted upon this knowledge. On average, following right/wrong feedback, participants spent 3.5 seconds reviewing each sentence corresponding to an error made on the initial test, as compared to 3.7 seconds on sentences containing correctly recalled facts. Participants did show some sensitivity in their review behaviour, spending more time reviewing previously tested facts than not tested facts (a reasonable strategy, given that in real life things that are tested on an initial test often appear on later tests). However, they did not discriminate within previously tested facts.

One issue may have been that the right/wrong feedback was relevant to only a fraction of the to-be-reviewed information. Participants reviewed 172 sentences, only 24 of which were relevant to the initial test, and of those only 8 were paired with right/wrong feedback. In other words, reviewing involved shifting through a large amount of information, most of which had not been associated with any feedback at all. Reviewing did greatly improve performance on the final test—but having received right/wrong feedback did not change or facilitate review behaviour.

To better test the hypothesis about right/wrong feedback and review, we made several changes in the procedure for Experiment 2. First, feedback condition was manipulated between participants, to increase the number of observations associated with right/wrong feedback. We also changed the materials so that everything in the review period would be relevant. Because this was difficult to do with the passages (which contained many filler sentences without facts), we switched to the Luganda–English word pairs previously used by Pashler and colleagues (2005). These changes in design and materials also had the advantage of making the no review conditions of Experiment 2 essentially replications of Pashler et al., to which we compared our review conditions.

Finally, we added the collection of confidence ratings to both the initial and final tests. This change was motivated by recent work by Butler and colleagues (Butler et al., 2008). As described briefly in the introduction, Butler and colleagues found that answer feedback helped low-confidence correct answers to persist onto the final test (in addition to supporting error correction). Without

feedback, only about 40% of the low-confidence correct answers from the initial test were produced on the final test. With answer feedback, however, 80% of questions initially answered correctly with low confidence were again answered correctly on the final test. Thus, answer feedback improved retention of initially correct answers, albeit low-confidence ones. Our interest was whether right/wrong feedback would have a similar benefit for low-confidence correct answers.

In short, in Experiment 2 participants studied Luganda–English word pairs and then took an initial translation test. One group of participants received answer feedback about their translations, another right/wrong feedback, and the third group no feedback. Half of the participants then had a chance to review the original study list while the others did an unrelated task, and all participants then took a final translation test. Both translation tests required participants to judge their confidence in each answer.

EXPERIMENT 2

Method

Participants. A total of 83 Duke University undergraduates participated for pay or to partially satisfy a course requirement. Participants were tested individually or in small groups of up to three people. Eight participants were excluded due to extremely poor or excellent performance on the initial test (5% correct or less, or 80% correct or greater; no participant in Experiment 1 met these exclusion criteria). Three additional participants were excluded because reaction times were not recorded. Therefore the data from 72 participants were analysed, with 24 in each of the three feedback conditions.

Design. The design was a 3 (feedback: none, right/wrong, answer) $\times$ 2 (review: none, review) between-participants design. As in Experiment 1, the main dependent measure was the change in performance from Test 1 to Test 2. We also examined the mean time spent reviewing the word pairs.

Materials. We chose 20 English–Luganda word pairs from Pashler et al.'s (2005) materials. The items were selected for ease of spelling in English, to minimise misspellings (and thus increasing the accuracy of feedback in the right/wrong condition). The Luganda materials used in Experiments 2 and 3 are available from the corresponding author on request.

Both initial and final tests were cued recall tests: each Luganda word was presented and the participant typed in the English translation. The words were tested in a random order using DirectRT software (Jarvis, 2004). The word pairs, feedback, and test items were presented in size 38 Times New Roman white font on a black computer screen.

Procedure. Like Experiment 1, the procedure of Experiment 2 consisted of five main parts: a study period, an initial test, a review period, a filled delay, and a final test. The procedure of Experiment 2 paralleled Experiment 1 except for some changes (such as the switch to a between-participants manipulation of feedback condition) to better parallel Pashler et al. (2005).

During *study*, participants were instructed to learn Luganda–English word pairs. Participants read an example Luganda–English word pair before beginning the study phase, which involved the presentation of the 20 critical word pairs in a random order. Each Luganda–English word pair was presented once for 6 seconds with a 2-second inter-stimulus interval.

Immediately after the study phase, participants began the second part of the experiment, the *initial test*. As in Experiment 1, they read instructions encouraging them to check their answers for misspellings to avoid errors in scoring. They also answered a practice question before beginning the test. All 20 studied word pairs were tested in a random order. The Luganda word was displayed and the participant was instructed to type the English translation. For example, "*Leero means*:" appeared on the screen and the participant typed their translation into a text box. Immediately following each response, participants rated their confidence from 1 (sure wrong) to 5 (sure correct). The feedback conditions were the same as in Experiment 1, except that in Experiment 2 feedback was manipulated between participants. Participants in the no feedback condition received no feedback and simply proceeded to the next question after rating their confidence. Depending on the correctness of each response, participants in the right/wrong feedback condition saw the word "correct" or "incorrect" for 5 seconds before the computer advanced to the next question. In the answer feedback condition, the correct English translation was displayed for

5 seconds before the computer advanced to the next question.

The *review period* began with a 40-second visuo-spatial puzzle, to separate the initial test from the review. Next, participants in the review condition viewed each word pair for as long as they wanted. To proceed to the next word pair, participants hit the spacebar, and 2 seconds later the next word appeared. Instead of reviewing, participants in the no review condition spent 80 seconds[3] solving more two visuo-spatial puzzles.

Following the review period, during the *filled delay* all participants spent 40 seconds solving an additional visuo-spatial puzzle. This delay was shorter than that used in Experiment 1 as we expected our participants to find Luganda words harder to remember than general knowledge facts.

Lastly, participants proceeded to the *final test*. The final test was identical to the initial test except that it never included feedback. For each Luganda word, participants typed the English translation and rated their confidence in their answer. The same instructions were given as for Experiment 1: participants were warned not to guess and to type ''I don't know'' if they did not know the answer.

Following the final test, participants were debriefed and thanked for their participation.

Results

All results are significant at the .05 level unless otherwise noted.

Change in proportion correct from the initial to the final test. The middle portion of Table 1 shows the proportion of questions answered correctly on the initial and final tests, as well as the change across tests. As in Experiment 1, there was some variability in initial test performance (even though the feedback and review manipulations had not yet occurred), but the conclusions were the same when the analyses were computed on proportion correct (and time of test was used as a within-participants factor).

Overall, answer feedback ($M = .34$) was more effective than both right/wrong feedback ($M = .22$) and no feedback ($M = .19$), $F(2, 66) = 7.53$, $MSE = .02$ (to see this main effect in Table 1, the reader should collapse across the two review conditions). In addition, the review period was helpful, $F(2, 66) = 117.24$, $MSE = .02$. Participants who were able to review the word pairs, regardless of feedback condition, ($M = .43$), showed greater improvement on the final test than did participants who did not review the words ($M = .07$).

There was no interaction between type of feedback and review, $F < 1$. In contrast to Experiment 1 (where review was only helpful following right/wrong or no feedback), in Experiment 2 reviewing helped regardless of feedback condition.

Effects of feedback and review on initially correct answers. A series of conditional analyses again allowed a closer examination of the data. We began with an analysis of Test 2 scores conditional upon having been answered *correctly* on the initial test. As shown in the middle of Table 2, neither feedback nor review had any effect for items that were already correct. Replicating Pashler et al. (2005) and Experiment 1, when participants already knew the correct answer, performance on the final test was always high regardless of feedback condition, $F < 1$. Participants translated numerically more Luganda words correctly following review ($M = .95$) than if they had not been able to review the material ($M = .90$), but this effect was not significant, $F(1, 66) = 2.00$, $p = .16$. There was no interaction between feedback and review, $F(2, 66) = 1.05$, $p = .36$. In short, similar to the first experiment, there was no strong evidence for benefits of review or feedback for items answered correctly on the initial test.

It was our intention to also examine correct answers on the initial test split by confidence levels. We were interested in whether feedback affected the persistence of low-confidence correct answers, as found by Butler et al. (2008). However, because very few participants produced correct answers with low levels of confidence, we did not have enough observations to split the data by confidence level. Even when the three lowest confidence levels were collapsed, there were only five participants included in each of the three feedback conditions, and on average they each contributed only 1.40 observations.

Effects of feedback and review on initial errors. Next we examined final test performance for items answered *erroneously* on the initial test.

[3] The filler task was shorter for no review participants in Experiment 2 than in Experiment 1, as it was expected that participants in Experiment 2 would need less time to review 20 word-pairs than was needed to review 172 sentences in the first study.

These data can also be found in the middle of Table 2. Three participants were not included because they did not erroneously translate any words on the initial test. Note this does not mean that the participants correctly translated all of the items, rather they often responded with "I don't know".

Receiving answer feedback after errors led to more correct answers on the final test ($M = .70$) than did right/wrong feedback ($M = .45$) or no feedback ($M = .42$), $F(2, 63) = 7.34$, $MSE = .07$. In addition, participants who reviewed the material successfully translated more Luganda words ($M = .79$) than did participants who were unable to review ($M = .26$), $F(1, 63) = 68.82$, $MSE = .07$. Reviewing the material helped participants in all three feedback conditions, but without a review period only the answer feedback group showed improvement. This resulted in an interaction between feedback and review, $F(2, 63) = 3.99$, $MSE = .07$.

Review time. The review times for Experiment 2 can be seen in the bottom portion of Table 3. One participant was missing data from one or more cells, and therefore was not included in Table 3 or in the ANOVA.

Unlike in Experiment 1, participants in this experiment showed metacognitive awareness and changed their review activity based on the correctness of their answers. If an item had been correctly translated on the initial test, participants spent less time reviewing it ($M = 3.12$) than words previously translated erroneously ($M = 8.60$), $F(1, 41) = 61.78$, $MSE = 10.67$. However, the type of feedback received had no effect on later study times. Participants in all feedback conditions spent equal time reviewing the items ($M =$ 5.86 s), $F < 1$ and there was no interaction between feedback condition and review, $F < 1$.

Discussion

The results in the no review condition were as expected: when participants were not able to review the material between tests, only answer feedback led to an improvement on the final test. The benefits of answer feedback were specific to items answered erroneously on the initial test: neither answer nor right/wrong feedback had an effect on words initially translated correctly. These results replicate those of Pashler and colleagues (2005) and our Experiment 1.

Review was again a powerful intervention, improving performance on the second test in all conditions. The no feedback and right/wrong feedback conditions showed identical effects of review: both showed similar levels of improvement (about 40% on average) when participants were allowed to review the material. Having received right/wrong feedback did not make review more powerful: participants in the no feedback condition benefited just as much from review as did participants in the right/wrong feedback condition. The reaction time data in the review condition help to explain this result. In both the no feedback and the right/wrong feedback conditions, participants spent less time studying word pairs that they had correctly translated on the initial test. In other words, the review times of the no feedback participants suggest that these participants knew which words they had translated correctly (and which ones they had translated erroneously)—meaning that the right/wrong feedback would not have provided any additional information. On the one hand these data differ from what was observed in Experiment 1, where participants reviewed facts for similar amounts of time regardless of whether or not they were retrieved on the initial test. But in both experiments the bottom line about right/wrong feedback was the same: having received right/wrong feedback did not make review any more effective than it was following no feedback. The switch in materials from prose to vocabulary meant that participants were better able to judge what they did versus did not know, and adjust their review accordingly, but it did not change conclusions about right/wrong feedback.

There was one other difference in the results across the two experiments, which can be seen in Table 1. The feedback by review interaction was significant in Experiment 1 but not Experiment 2. To be clear, the conclusions about right/wrong versus no feedback were the same across the two experiments: both benefited from review, to similar extents. The only difference across the experiments is in the answer condition: for some reason, there was no benefit from reviewing following answer feedback in Experiment 1. While somewhat puzzling, this result does not change our conclusions about right/wrong feedback. The conclusions across the two experiments were the same if the more sensitive conditional analyses are considered; for error correction, the interaction was significant in both experiments.

Finally, we had hoped to examine low-confidence correct answers, to see if participants who received right/wrong feedback would be more likely to repeat their initial correct guesses on the final test. Unfortunately we did not have enough observations to perform the analysis. Therefore Experiment 3 focused on this analysis and did not include a manipulation of review (since the conclusions about review following right/wrong feedback were similar across the first two experiments). In addition, we made a number of changes to ensure that there would be enough low-confidence correct answers to analyse. First, we increased the number of word pairs from 20 to 40, increasing the number of observations. Second, we decreased our confidence scale from five points to the 4-point scale used by Butler and colleagues (2008). Third, the experiment began with a familiarisation phase during which the participants learned the 40 English words that would appear in the later translations. Finally, participants were forced to give an answer to each question on the initial test; Butler et al. (2008) identified forced responding as critical for obtaining low-confidence correct answers. The participants were instructed never to respond "I don't know", and instead were told to guess one of the words learned during the familiarisation phase.

EXPERIMENT 3

Method

Participants. A total of 102 Duke University undergraduates participated to partially satisfy a course requirement. Participants were tested individually or in small groups of up to four people. Ten participants were excluded due to extremely poor or excellent performance on the initial test (as in the earlier experiments, 5% correct or less, or 80% correct or greater). Therefore, the data from 92 participants will be presented here, with 32 participants in the no feedback condition, 30 in the right/wrong feedback condition, and 30 in the answer feedback condition.

Design. The experiment had a single between-participants manipulation of feedback condition with three levels: no feedback, right/wrong, and answer. The main dependent measure was change in performance from the initial to the final test. Of particular interest was how any change across tests differed as a function of confidence in Test 1 answers.

Materials. The materials consisted of the 20 Luganda–English word pairs used in Experiment 2, plus 20 additional Luganda–English word pairs. Just like the first 20, the 20 new items were selected for ease of spelling in English in order to minimise misspellings.

As in Experiment 2, both the initial and final tests were cued recall tests in which Luganda words were presented and the participants typed in the English translations. The word pairs, feedback, and test items were presented just as they were in Experiment 2.

Procedure. The experiment included five main parts: a familiarisation phase in which participants learned the English words that would be paired with the Luganda words, a study phase, an initial test, a filled delay, and a final test. In contrast to the first two experiments, there was no review phase in Experiment 3.

The *familiarisation phase* was instituted to elicit low-confidence correct answers on the initial test. The familiarisation phase consisted of two study-test trials to teach participants the English words that they would have to produce during the translation tests. During this familiarisation phase, participants studied the English words and never saw any Luganda words. On the first study trial, participants categorised each of the 40 English words as concrete or abstract. The English words appeared one at a time in random order, and each remained on the screen until a judgement was entered. Participants then recalled as many of the English words as possible. The second trial was identical to the first, except that participants were asked to categorise each word as pleasant or unpleasant, and then they recalled the list of words a second time.

The *study* phase was the same as in Experiment 2, except that 40 word pairs were presented. Following study, all participants completed the *initial test*, a cued recall task that consisted of all 40 Luganda words. This test differed from that used in Experiment 2 in one major way, beyond the increase to 40 items. Critically, the instructions were changed to warn against skipping items or answering with "I don't know". Rather, if they did not know the answer to a question, participants were instructed to type in their best guess from the list of English words that they studied earlier. These instructions were designed to increase the number of low-confidence correct

responses. As in Experiment 2, participants rated their confidence in each translation. A "1" indicated a guess, a "2" low confidence, a "3" medium confidence, and a "4" high confidence that the answer was correct. Feedback after each question was given just as in Experiment 2.

participants next completed the *filled delay*, which consisted of 160 seconds of visuo-spatial puzzles.

The fifth phase was the *final test*. This final test was the same as the one used in Experiment 2, except for the increased number of translations. No feedback was given, and participants were instructed *not* to guess (but rather to type "I don't know"). The instructions explicitly noted that the forced responding instructions used for the initial test no longer applied. Participants were then debriefed and thanked for their participation.

Results

All results are significant at the .05 level unless otherwise noted.

Familiarisation phase. We first examined the number of English words that the participants were able to recall during the familiarisation phase. After two study-test trials, participants were able to recall almost half of the words. As expected, participants recalled more words on the second test ($M = .49$) than on the first test ($M = .32$), $F(1, 89) = 275.77$, $MSE = .01$, and as expected there was no effect of feedback condition, $F < 1$.

These results show that participants knew enough of the English words to guess translations on the initial translation test, and in fact 91% of errors on the initial test were studied English words. The familiarisation phase successfully led participants to follow the instructions on the initial test.

Change in proportion correct from the initial to the final test. We again examined change in performance from the initial test to the final test as a function of feedback condition. We subtracted the initial test score from the final test score, and this change score was the dependent variable in the ANOVA (and the same results were obtained when time was treated as an independent factor in the ANOVA). The relevant data are shown in the bottom portion of Table 1.

As in the no review conditions of Experiments 1 and 2, only answer feedback led to improved performance across tests, yielding a main effect of feedback condition $F(2, 89) = 69.74$, $MSE = .003$. Participants who received answer feedback improved from 30% correct on the initial test to 42% on the final test, whereas participants who received right/wrong or no feedback dropped an average of 3% across tests.

Effects of feedback on initially correct answers. Given that a word was *correctly* translated on the first test, how likely was it to be correctly translated on the final test? The relevant data are shown in the bottom panel of Table 2. As in the first two experiments, correct answers were likely to persist to the final test ($M = .79$), and this was unaffected by feedback condition, $F < 1$.

More critical for present purposes is whether the persistence of correct answers depended on confidence in one's initial response. This analysis requires participants to have made low-confidence correct responses as well as high-confidence ones (which did not occur in Experiment 2). Table 5 shows that the addition of the familiarisation phase increased low-confidence correct responses. Given that a range of confidence was observed, we examined whether feedback helped low-confidence correct answers persist to the final test. The full set of data appears in Figure 1, and the number of participants contributing to each point can be found in Row 1 of Table 5. To deal with this missing data problem we collapsed the two lower confidence ratings (responses of 1 or 2) into a single "lower confidence" category, and the two higher

TABLE 5
Distribution of lower- and higher-confidence correct responses on the initial test, from Experiment 3

	1 Guess	*2 Low*	*3 Medium*	*4 High*	*1 or 2 Lower*	*3 or 4 Higher*
Number of participants with at least one observation (out of 92)	53	48	68	91	75	92
Average number responses (participants with at least *one* observation)	1.85	1.44	2.22	7.79	2.23	9.35

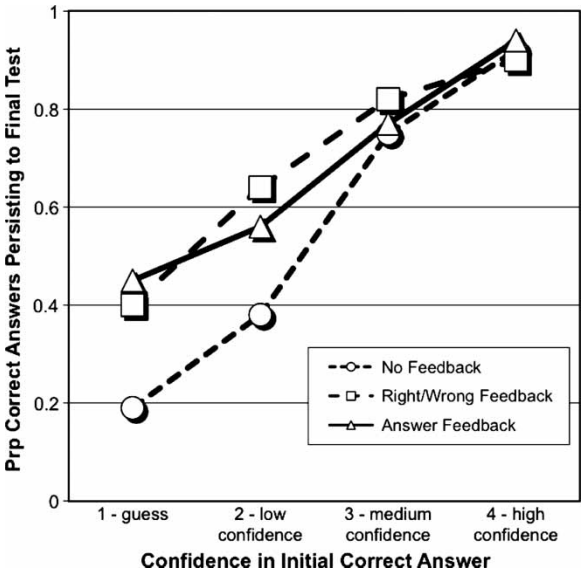

Figure 1. The proportion of correct answers from the initial test that persisted to the final test, as a function of confidence in the Test 1 response and feedback condition (Experiment 3).

confidence ratings (responses of 3 or 4) into a single ''higher confidence'' category (the right portion of Table 5 contains information about the number of resulting observations). We then carried out a 3 (feedback condition: none, right/wrong, answer) $\times$ 2 (confidence: lower, higher) ANOVA on the proportion of initially correct items that were answered correctly on the final test. This analysis excluded 17 participants because they made no lower confidence correct answers on the initial test. Of the 75 participants included in the analysis, 25 received no feedback, 22 right/wrong feedback, and 28 answer feedback.

As expected, there was a large effect of initial confidence. Higher-confidence correct answers were more likely to persist to the final test ($M = .90$) than were lower-confidence correct answers ($M = .44$), $F(1, 72) = 103.51$, $MSE = .08$. There was no overall effect of feedback condition, $F(1, 72) = 1.71$, $p = .19$, but the predicted interaction between feedback condition and confidence level was significant, $F(2, 69) = 3.16$, $MSE = .08$. Persistence rates were high for initially correct answers made with higher confidence, and this did not differ as a function of feedback condition (.92 for no feedback, .88 for right/wrong feedback and .91 for answer feedback). In contrast, for correct answers initially made with lower confidence, both right/wrong ($M = .52$) and answer ($M = .50$) feedback were more effective than no feedback ($M = .30$), $t(45) = 2.02$, $SE = .11$, $p = .05$; $t(51) = 1.96$, $SE = .10$, $p = .06$, respectively.

Effects of feedback on initial errors. Although not the central focus of Experiment 3, for completeness we also report the effects of feedback given an *error* was made on the initial test. For items that were initially answered erroneously, only answer feedback led to an improvement in performance across tests ($M = .28$), $F(2, 89) = 61.41$, $MSE = .01$. Very few errors were corrected in the No Feedback or Right/Wrong Feedback conditions (both Ms $= .03$). For purposes of contrasting these means with those obtained in the earlier experiments, they can be found in the bottom row of Table 2.

Discussion

As in Experiments 1 and 2, only answer feedback was effective in increasing the proportion of correct answers from the initial test to the final test, and in correcting initial errors. In Experiment 3, however, participants who received right/wrong feedback did show one advantage over participants who received no feedback. Low-confidence correct answers were more likely to persist onto the final test after right/wrong feedback than after no feedback. The data replicate those of Butler and colleagues (2008) and extend their finding from answer to right/wrong feedback.

To understand why feedback stabilises access to low-confidence correct answers, we can link to the literature on the *hypercorrection effect.* The hypercorrection effect is the finding that high-confidence *errors* are more likely to be corrected following feedback than low-confidence errors (Butterfield & Metcalfe, 2001). One explanation for the hypercorrection effect is that participants pay more attention to the feedback when it is surprising. Learning that a confidently made response is wrong is surprising, and thus one pays more attention to that feedback, with consequences for later memory. A similar argument has been made for low-confidence correct answers: participants are surprised when their guesses are correct and thus better remember that feedback too. The result is not a hypercorrection effect per se (since there is nothing to correct when participants guess the right answer), but the mechanism is similar: people pay more attention to stimuli when they find them surprising. The

results are consistent with other data involving source judgements: participants are better at remembering the appearance of feedback associated with both high-confidence errors and low-confidence correct responses (Fazio & Marsh, 2009; see also Butterfield & Metcalfe, 2006).

GENERAL DISCUSSION

We examined three ways that receiving right/wrong feedback might be expected to improve performance over that observed in a no feedback condition. In only one of the three situations was right/wrong feedback effective: right/wrong feedback led to greater persistence of low-confidence correct answers in Experiment 3. For the retention of low-confidence correct answers, right/wrong feedback was just as effective as answer feedback. This finding supports the argument that there is some value to the information that right/wrong feedback conveys (namely, that something is correct).

However, overall, right/wrong feedback was not very effective. The data from the no review condition of Experiment 1 replicated the results of Pashler et al. (2005) with Luganda–English word pairs and extended them to facts embedded in prose passages. When participants did not have a chance to review the material, answer feedback was the only type of feedback that led to improvement on the final test. Even more important was the finding that the combination of right/wrong feedback and review was not helpful. Originally we believed that right/wrong feedback would be useful if participants had a chance to encounter the answer feedback later, and in two experiments we tested this by giving some participants a chance to review the original study materials. Participants in Experiments 1 and 2 did benefit from reviewing the original materials, but review was equally useful regardless of whether participants had previously received right/wrong feedback or no feedback at all. The review times tell the same story: compared to the no feedback condition, having received right/wrong feedback did not lead to disproportionate attention to items previously answered erroneously.

The data were clear that reviewing material was an effective way of providing feedback, and in that way they are consistent with other data supporting the value of re-reading text (e.g., Bromage & Mayer, 1986; Haenggi & Perfetti, 1992). This was especially true when the materials consisted of Luganda–English words pairs (in Experiment 2); in this case, participants spent far more time attending to words they had not been able to translate on the initial test.

Our data support prior work suggesting that if feedback is to be provided it should include the correct answer, as opposed to simply marking a response as correct or erroneous. It is unlikely that a student will make enough low-confidence correct answers to make right/wrong feedback useful in an everyday setting. Our data also support that review is an effective way of delivering feedback, at least for the types of materials considered here. From the teacher's perspective review is relatively more efficient; it is easier (and takes less time) for the teacher to the tell students to ''find the answers'' than to provide them. But of course, from the student's perspective, even if review is an effective way of obtaining feedback, it is less efficient—it will take the student more time to review and find the answers than to receive answer feedback. If students have a limited amount of time (and reviewing would take them away from some other meaningful activity), then it will probably be best to give answer feedback.

Manuscript received 8 June 2009
Manuscript accepted 12 January 2010

REFERENCES

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bromage, B. K., & Mayer, R. E. (1986). Quantitative and qualitative effects of repetition on learning from technical text. *Journal of Educational Psychology*, *78*, 271–278.

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

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

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

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627–634.

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

Glover, J. A. (1989). The 'testing' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

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

Haenggi, D., & Perfetti, C. A. (1992). Individual difference in reprocessing of text. *Journal of Educational Psychology*, *84*, 182–192.

Jarvis, B. G. (2004). *DirectRT (Version 2004.1.0.55)* [computer software]. New York: Empirisoft Corporation.

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

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*, 596–606.

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.

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

Roediger, H. L. III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 1155–1159.

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

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Swindell, L. K., & Walls, W. F. (1993). Response confidence and the delay retention effect. *Contemporary Educational Psychology*, *18*, 363–375.

Waldrop, P. B., Justen, J. E. III, & Adams, T. M. II. (1986). A comparison of three types of feedback in a computer-assisted instruction task. *Educational Technology*, *26*, 43–45.

Wentling, T. L. (1973). Mastery versus nonmastery instruction with varying test item feedback treatments. *Journal of Educational Psychology*, *65*, 50–58.