Mem Cogn (2012) 40:514–527
DOI 10.3758/s13421-011-0167-z

Making related errors facilitates learning, but learners do not know it

Barbie J. Huelser · Janet Metcalfe

Published online: 9 December 2011

Abstract Producing an error, so long as it is followed by corrective feedback, has been shown to result in better retention of the correct answers than does simply studying the correct answers from the outset. The reasons for this surprising finding, however, have not been investigated. Our hypothesis was that the effect might occur only when the errors produced were related to the targeted correct response. In Experiment 1, participants studied either related or unrelated word pairs, manipulated between participants. Participants either were given the cue and target to study for 5 or 10 s or generated an error in response to the cue for the first 5 s before receiving the correct answer for the final 5 s. When the cues and targets were related, error-generation led to the highest correct retention. However, consistent with the hypothesis, no benefit was derived from generating an error when the cue and target were unrelated. Latent semantic analysis revealed that the errors generated in the related condition were related to the target, whereas they were not related to the target in the unrelated condition. Experiment 2 replicated these findings in a within-participants design. We found, additionally, that people did not know that generating an error enhanced memory, even after they had just completed the task that produced substantial benefits.

Keywords Memory · Errors · Generation · Metacognition · Associative learning

B. J. Huelser (✉) · J. Metcalfe
Department of Psychology, Columbia University,
401 Schermerhorn Hall, 1190 Amsterdam Ave., MC 5501,
New York, NY 10027, USA
e-mail: bjh2135@columbia.edu

This article addresses the effect of making errors on learning. Should one learn by studying materials without making mistakes or by attempting to produce answers and committing inevitable errors that such attempts entail? When errors are left uncorrected, they typically remain incorrect (Butler, Karpicke & Roediger, 2008; Fazio, Huelser, Johnson & Marsh, 2010; Metcalfe & Kornell, 2007; Pashler, Cepeda, Wixted & Rohrer, 2005; Pashler, Zarow & Triplett, 2003). However, feedback is highly effective in allowing the learner to correct previously incorrect answers (Butler et al., 2008; Metcalfe, Kornell & Finn 2009; Pashler et al., 2005; Pashler et al., 2003). In this article, only errors followed by corrective feedback were considered. The question here was whether, and under what conditions, committing an error facilitates learning. Although the main focus of this article is the memorial consequences for errorful, as compared with errorless, learning, a related question of interest is the following: Are learners *aware* of the circumstances in which committing errors can be effective for improving learning? Accurate metacognitive knowledge is important for metacognitive control and strategy selection (Kornell & Son, 2009; Metcalfe & Finn, 2008). If the learner is not aware of the potential efficacy of a learning strategy, he or she might implement suboptimal strategies. Hence, people's metacognitions about the effects of errors may be nearly as important as the effects of the errors themselves.

From a theoretical standpoint, there is reason to believe that even corrected errors might impede learning. An error, in essence, is often thought to be conflicting or competing information with regard to the correct response. As such, it should create an interference situation. In standard proactive interference paradigms, the first pairing of a target (B) with a particular cue (A) results in interference when cue A is later paired with a different response (C) (J. R. Anderson & Reder, 1999; M. C. Anderson & Neely, 1996; Barnes & Underwood, 1959; Loftus, 1979; McGeoch, 1942; Melton & Irwin, 1940; Osgood, 1949; Webb, 1917). Although there are several theories concerning how this interference arises (e.g., J. A. Anderson, 1973; J. R. Anderson & Bower, 1972; Eich, 1982; Gillund & Shiffrin, 1984; Hintzman, 1984; Metcalfe, 1990; Osgood, 1949), there is general agreement that it does occur. Interference from errors might be expected to be even greater than interference theory would normally predict, since interference theory does not take into account whether or not the interfering information is self-produced. Incorrect information that is self-generated might be even more difficult to overcome than a provided response, because the process of self-generation has been shown to enhance memory for the response (Slamecka & Graf, 1978; for reviews, see Bertsch, Pesta, Wiscott & McDaniel, 2007; Mulligan & Lozito, 2005).

In accordance with the rationale described above, it has sometimes been recommended that errors be eliminated during learning (Glaser, 1990). For example, Guthrie (1952) suggested that errors should be avoided because, when errors are practiced, the incorrect response to a particular stimulus will be strengthened. Furthermore, errorless, as compared with trial-and-error, learning has been shown to be beneficial for people with memory impairments, including Alzheimer's disease, schizophrenia, Korsokoff's syndrome, and trauma (see Clare & Jones, 2008, for a review). One concern with generalizing from this line of empirical research, however, is that the benefits of errorless over errorful learning have been found primarily in patient populations and may not apply to typical learners. Nevertheless, in an experiment by Cunningham and Anderson (1968), worse retention was found after typical participants had been forced to guess, rather than following a simple presentation of the to-be-remembered material.

Despite the arguments that the generation of errors impedes learning, several researchers have found that error-generation is not detrimental to memory of subsequently learned correct answers. One way of examining the effect of errors on learning is by forcing responses for every item on a test, as compared with leaving participants free to answer only when they so choose. Forced responding results in more errors than does free responding. However, on a later test of definition terms, using this procedure with both college undergraduates and 6th grade students, Metcalfe and Kornell (2007) found neither benefit nor impairment for forced, as compared with free, responding. Similarly, Kang et al. (2011) found that forced guessing did not lead to better or worse memory for the correct answer on a later retention test, neither immediately nor at a 1-week delay. However, it is impossible to know whether the lack of a difference might have occurred because people in the free-responding condition generated errors to the same extent as people in the forced-responding condition, but did not overtly express them. It is also not known what kinds of errors were produced under the forced-guessing procedures and, in particular, whether they were related or unrelated to the targets. Research based on multiple-choice quizzing prior to learning a lesson in a classroom setting also suggests that pretesting—which results in many errors—neither helps nor hurts memory for the correct information (McDaniel, Agarwal, Huelser, McDermott & Roediger, 2011). No difference in memory was found for items quizzed on a pretest, as compared with nonquizzed items.

In contrast to the findings above, however, there are some studies showing that under certain circumstances, making errors helps learning. Richland, Kornell and Kao (2009) found enhanced memory for material from reading passages when the to-be-remembered material was tested using cued recall questions prior to the reading of the passages, even though participants did not answer these pretest questions correctly. Izawa (1967, 1970) has also shown that multiple incorrect retrieval attempts enhanced learning: Producing more incorrect responses before receiving feedback led to better memory for the correct feedback than did producing fewer incorrect responses. Parlow and Berlyne (1971) found that participants were better at learning the correct translations for foreign language words when they had previously made an erroneous guess, as compared with when they were exposed to the guesses of others. Kane and Anderson (1978) showed that generating the last word of the sentence, even if it was incorrect, led to enhanced performance over simply reading the sentence. Slamecka and Fevreiski (1983) reported a benefit, above just reading the answer, from trying unsuccessfully to generate it.

Finally, in a paradigm that we will investigate here, Kornell, Hays and Bjork (2009) demonstrated a considerable benefit of prior incorrect guessing for subsequent learning of the correct answer. Participants learned weakly associated word pairs (e.g., *whale–mammal*, *swing–tree*, *together–love*) for a later cued recall test. During the initial learning phase, participants randomly studied word pairs in either a reading mode or an error-generating mode. In the reading mode, both the cue and the target were displayed on the screen for a fixed amount of time (either 5 or 13 s). In the error-generating mode, participants saw only the first word (the cue) for 8 s and had to type a guess into the computer as to what they thought the target would be, followed by the correct cue–target pairing displayed for 5 s. At test, given the cue, participants were required to produce the correct target and not the original error. Error-generation led to enhanced retention, as compared with both reading conditions.

In sum, it is unclear whether errors during learning hinder, enhance, or simply have no effect on learning. Any of these three options might be possible under different conditions, but it is not yet known what those conditions might be. However, studies in which there was a benefit of error-generation used cue–target pairs that generally seemed to be meaningfully related. For example, the experiments in Kornell et al.'s (2009) study showing beneficial effects used to-be-remembered materials that were weakly associated word pairs. By extension, it might be plausible that errors generated in response to these cues might also have been related, rather than unrelated, to the targets. However, in one of Kornell et al.'s (2009) experiments, no benefit for error-generation was found. In this case, participants guessed answers to fictional general knowledge questions (Berger, Hall & Bahrick, 1999) to which they could not possibly have known anything about the correct answers, such as “What is the last name of the person who invented maladaptibility?” It is likely that the errors that people generated in this particular case were unrelated to the targets. Additional support for the idea that the relatedness of the errors might matter comes from Slamecka and Fevreiski (1983), who compared a generation-followed-by-feedback condition with a read condition. Judges retrospectively evaluated the relatedness of the errors of commission that participants had made, dividing them into those that were related and unrelated to the target. Related errors led to fairly high later recall, whereas unrelated errors and omissions led to low recall. These results suggest that the relatedness of the errors to the target may be an important factor in determining whether errors help or hurt recall—a possibility that was investigated in the experiments that follow.

Finally, given that there is a conflict concerning the effects of errors in the research literature, it is plausible to suppose that the learners themselves might not know whether errors help or hurt learning. As well as exploring the conditions under which errors promote and hinder learning, we also investigated whether, in retrospect, participants were able to accurately monitor whether generating errors helped or hurt their performance on the final test. This question is important, since metacognitive monitoring has been shown to have consequences for strategy selection, referred to as *metacognitive control* (Metcalfe & Finn, 2008; Thiede, Anderson & Therriault, 2003).

Experiment 1

In Experiment 1, we extended Kornell et al.’s (2009) experiment by contrasting memory for weakly associated word pairs, for which they found the beneficial effect of error-generation, to unrelated word pairs, for which we hypothesized that the effect would not be found. Participants studied word pairs in an error-generation condition and two different read conditions. Half of the participants studied weakly related word pairs, while the other half studied unrelated word pairs. We also tested participants’ retrospective metacognitions about their memory performance.

Method

Participants Seventy-one Columbia undergraduates participated for partial fulfillment of a class requirement. Data from 11 nonnative speakers were excluded, leaving data from 60 participants. Mean age was 21.8 years (*SD* = 6.2), and 68.3% of the participants were female. All participants in both experiments were treated in accordance with APA ethical guidelines.

Design and materials The semantic relation of the to-be-remembered materials was manipulated between participants, while learning condition was a within-participants variable, resulting in a 2 (materials: related, unrelated) × 3 (learning condition: read-short, read-long, error-generate) mixed design.[1]

For the related materials condition, 90 weakly associated word pairs were selected from Nelson, McEvoy and Schreiber (1998) norms, closely following Kornell et al.’s (2009) word pair selection criteria. Given the first word, approximately 5% of participants in Nelson et al.'s experiment produced the target as the first associate. Specifically, forward associative strength was between .05 and .054, and backward associative strength was 0. Each word was a minimum of four letters long. For the unrelated materials condition, new materials were selected because in a pilot experiment, cued recall performance was at floor for random word pairs created from the Nelson et al. (1998) norms. Therefore, unrelated word pairs were created from Pavio, Yuille and Madigan (1968) norms. One hundred eighty words were selected (to create 90 word pairs) with relatively high concreteness ratings (6.38–7 on a 1–7 scale) and were a minimum of four letters long. Words were randomly assigned as cues or targets, and three independent coders checked that the so-constructed list of 90 unrelated word pairs contained no accidentally related word pairs.

[1] Because we did not know whether our experiment would replicate the findings of Kornell et al. (2009), we assigned the related materials condition to the first 18 participants, a condition that is most similar to their experiment. After the first set of data on 18 participants was collected in 3 days and it was clear that we were replicating the earlier results, we randomly assigned participants to both materials conditions, beginning the following week.

Mean concreteness ratings were the same for the words assigned as cues and targets (M = 6.77). For each of the between-participants conditions, the 90 word pairs were randomized into three sets of 30 items, which were rotated through each of the study conditions, creating three unique counterbalanced conditions.

Procedure This experiment had four phases: learning, distractor, final test, and metacognitive judgment. During the learning phase, 30 word pairs were presented in each of the three conditions (90 word pairs in total). Word pairs were presented in a random order by MediaLab and DirectRT software (Jarvis, 2004). In the error-generation condition, participants were given only the first word (cue) of a word pair, with a text box displayed below. Participants were instructed to think of what the second word might be and to type their response into the text box as quickly as possible. After 5 s, the text box disappeared, and the correct cue–target pairing appeared, with both the cue and the target remaining on the screen for 5 s. In the read-short condition, both the cue and the target were presented together on the screen for 5 s, while in the read-long condition, both the cue and target were presented for 10 s. These conditions were presented in a random order (not blocked). The computer made a soft clicking sound to alert the participants to the presentation of the next word pair. Before the study phase began, participants read instructions on a computer screen. The experimenter also discussed the procedure verbally and ensured that participants understood the task before proceeding. During the instructions, the experimenter expressed that it was extremely difficult to correctly guess the correct target word, to prevent the participants from being discouraged by poor performance on the task. They were instructed to remember the target answer presented by the computer for the later memory test, not the word they had produced. During the distractor phase, participants played a visuospatial computer game for 6 min before continuing to the final test.

The final test was self-paced and consisted of all 90 word pairs presented during the learning phase. For each word pair, the cue was displayed on the screen, with a textbox below. Participants were instructed to type in the correct target for each cue and to provide a guess if unsure of the correct answer. The order of presentation was randomized.

Following the final test, participants made a metacognitive judgment of their performance on the final test on the basis of the initial learning conditions. Instructions were as follows:

> There were three conditions in this experiment: A) together–short: both words displayed on the screen for 5 s, B) together –long: both words displayed on the screen for 10s; C) separate: the first word presented separately (5 s) before both words were displayed (5 s). Which condition helped you learn the word pairs the best for the final test? Please order the conditions in order from which condition led to the BEST to WORST memory on the final test.

Participants subjectively ranked the conditions by entering the associated letter from the best to the worst for memory. We avoided the word *error* in the error-generation condition because we thought that its negative connotation might bias the judgment. Following a demographic questionnaire, all participants were thanked and debriefed.

Results

Two coders checked for and corrected spelling and typographical mistakes on the original and final tests before analysis of the data. A strict coding rule was followed in which, if the tense (i.e., *clean* vs. *cleaned*/*cleaning*) or form of speech (*dust* vs. *dusty*) was different from the target, that item was coded as incorrect. However, in the few instances in which an item was made plural (*reptile* vs. *reptiles*), it was coded as correct.

Learning phase performance Participants in the related materials condition guessed correctly on 3% of the error-generation trials (SD = .03), while no participant in the unrelated materials condition ever correctly guessed the target word during the learning phase (M = .00, SD = .00). All further results reported for the error-generation condition are only from items that were initially answered incorrectly during the learning phase—that is, 97% of the trials for the related materials condition, and 100% of the trials for the unrelated materials condition.[2]

Final cued recall test correct performance As is shown in Fig. 1, correct final performance was higher for related materials (M = .64, SD = .19) than for unrelated materials (M = .21, SD = .15), $F(1, 58) = 91.34$, MSE = .09, $p < .001$, $\eta_p^2 = .61$. There was a main effect of learning condition: Error-generation led to the highest proportion correct on the cued recall test, $F(2, 116) = 13.71$, MSE = .01, $p < .001$, $\eta_p^2 = .19$. However, this main effect was qualified by an

[2] Of these errors, 90% were errors of commission for related materials, and 91% for unrelated materials, $t < 1$. Reported data include errors of omission as well, since correct performance on the final cued recall test was not statistically different as a function of prior error type, $F < 1$. For Experiment 2, 96% of errors were errors of commission, and a similar pattern of results for final test performance as a function of prior error type was found. Therefore, results are not conditionalized upon error type, with the exception of latent semantic analysis as it could be computed only for generated errors.

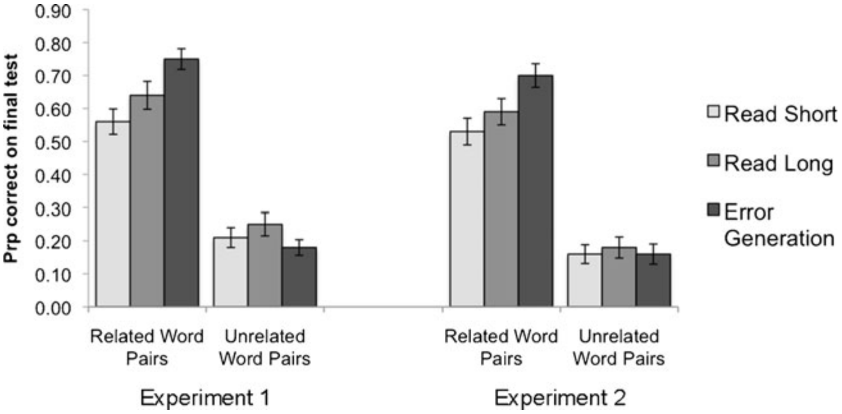

Fig. 1 Cued Recall Performance. Correct performance on final cued recall test as a function of Learning condition and Materials for both Experiment 1 (between-subjects) and Experiment 2 (within-subjects)

interaction with type of materials. Although error-generation enhanced retention for related materials, it did not enhance performance for unrelated materials, *F*(2, 116) = 32.21, *MSE* = .01, $p < .001$, $\eta_p^2 = .36$. Within related materials, the error-generation condition led to the highest proportion correct on the cued recall test (*M* = .74, *SD* = .17), which was much higher than recall in the read-long condition (*M* = .62, *SD* = .23), *t*(29) = 5.14, *SE* = .02, $p < .001$. The read-short condition led to the lowest proportion correct (*M* = .54, *SD* = .21), which was significantly lower than performance in the read-long condition, *t*(29) = 3.54, *SE* = .02, $p < .01$, and the error-generation condition, *t*(29) = 7.37, *SE* = .03, $p < .001$. With unrelated items, however, the read-long condition led to the highest correct performance (*M* = .25, *SD* = .19), which was significantly better than performance for both the read-short condition (*M* = .21, *SD* = .16), *t*(29) 2.09, *SE* = .02, $p < .05$, and the error-generation condition (*M* = .17, *SD* = .12), *t*(29) = 3.26, *SE* = .02, $p < .01$. Although the trend favored the read-short condition over error-generation, performance between these two conditions was not significantly different from one another, *t*(29) = 1.89, *SE* = .02, *p* = .068.

Reaction times Reaction time (RT) data on the final test were analyzed as a function of accuracy on the final test (correct vs. incorrect), learning condition (read-short, read-long, error-generation), and materials (related, unrelated; see Table 1 for means). RT data are reported in the present section for completeness but will be discussed only in the General Discussion section. Several participants did not have data in all cells in the RT data, and as a result, the degrees of freedom in the analyses given below differ from those given in the basic data for this experiment.

Overall, correct responses (*M* = 4.44 s, *SD* = 1.13) were faster than incorrect responses (*M* = 8.63 s, *SD* = 4.06), *F* (1, 54) = 67.65, *MSE* = 21.84, $p < .001$, $\eta_p^2 = .56$. Collapsed over accuracy, participants were slowest to respond to items on which they had previously generated an error (*M* = 7.35 s, *SD* = 3.43), in comparison with the read-short items (*M* = 6.20 s, *SD* = 2.68) and read-long items (*M* = 6.04 s, *SD* = 2.40), *F*(2, 108) = 13.38, *MSE* = 4.20, $p < .001$, $\eta_p^2 = .20$. There was an interaction between accuracy and learning condition, whereby the difference in RT between items answered incorrectly and correctly on the final test was larger in the error-generation conditions than in the read conditions, *F*(2, 108) = 6.30, *MSE* = 3.79, $p < .01$, $\eta_p^2 = .10$.

Lastly, the relatedness of the materials did not result in differences in RTs. Response latencies were similar regardless of materials condition. There was no difference between related and unrelated materials, *F* = 1.06, $\eta_p^2 = .02$, and materials did not interact with any other factor.

Error persistence In the error-generation conditions, more of the initially incorrect responses intruded on the final test for unrelated materials (*M* = .20, *SD* = .20), as compared with related materials (*M* = .05, *SD* = .06), *t*(58) = 4.05, *SE* = .04, $p < .001$.

Metacognition Data from 52 participants were included in the metacognitive analyses: 26 from the related materials condition and 26 from the unrelated condition. Exclusions were due to participant failure to assign a distinct metacognitive ranking to each of the three learning conditions. In order to compare performance and metacognitive rankings for each participant, the three conditions were assigned a value on a 0 to 2 scale. The learning condition in which the participant performed best on the final test was assigned a 2; the condition in which he or she performed second best was assigned a 1; the worst was given a score of 0. The same assignment was done for individuals' metacognitive ratings of the three learning conditions.

Table 1 Mean reaction time in seconds (s) for responding on the final test as a function of Learning condition, Material condition, and Accuracy on the final cued recall test. Standard deviations are provided in parentheses

	Correct on Final Test			Incorrect on Final Test		
	Read-Short	Read-Long	Error-Generate	Read-Short	Read-Long	Error-Generate
Experiment 1						
Related	3.86 (1.01)	3.93 (0.78)	4.27 (0.92)	7.83 (4.79)	7.26 (3.43)	10.10 (5.26)
Unrelated	4.85 (2.15)	4.57 (1.09)	5.14 (1.99)	8.27 (4.06)	8.42 (4.26)	9.87 (5.78)
Experiment 2						
Related	3.58 (1.25)	3.77 (0.92)	4.19 (1.33)	6.72 (2.96)	6.93 (2.96)	8.00 (3.830)
Unrelated	4.81 (1.83)	3.82 (1.24)	4.07 (1.46)	6.66 (2.71)	6.44 (2.16)	8.25 (2.91)

As can be seen in Fig. 2, participants believed that they performed the best in the read-long condition. They also believed that they had done poorly in both the error-generation condition and read-short condition—regardless of whether the materials were related or unrelated pairs. For the unrelated materials, these metacognitive rankings were approximately correct. However, the participants' beliefs were radically wrong for the related materials: They failed to realize that generating errors greatly facilitated recall under this condition, even after having just experienced the enhanced test performance.

To assess this pattern statistically, metacognitive mean ranking was contrasted with performance mean rankings within each learning condition. These comparisons were done separately for each of the two materials conditions, using the Wilcoxon nonparametric test in lieu of the standard paired-samples t-test. Rankings for performance and metacognitive judgments (within materials condition) are not independent, so these contrasts could not be computed. First, for the items in the read-short condition for related materials, there was a trend for actual performance (M = 0.35, SD = 0.56) to be worse than subjectively reported (M = 0.65, SD = 0.70), z = 1.86, p = .06. For the read-long condition, the mean metacognitive ranking was higher (M = 1.50, SD = .65) than the actual performance ranking (M = 0.92, SD = 0.61), z = 2.78, p < .01. Most interesting, however, in the error-generation condition, participants mistakenly believed that their performance was very low (M = 0.85 SD = 0.88) when it was actually high (M = 1.73, SD = 0.55), z = 3.45, p < .01. Within unrelated materials, participants' retrospective metacognitive rankings were very close to actual performance rankings: There was no difference in mean subjective metacognitive ranking, as compared with actual performance rank, for any of the comparisons, zs < 1.16.

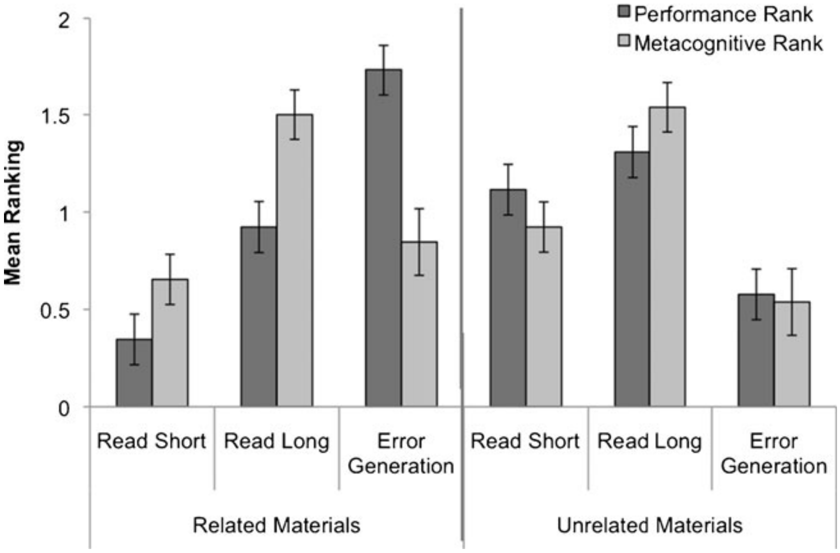

Fig. 2 Metacognitive Data for Experiment 1 (between-participants). Mean ranking of Learning conditions based on correct performance on the final cued recall test and subjective metacognitive judgments. The condition with the highest proportion correct or subjectively rated the best was assigned a score of a 2. Second best was assigned a score of 1 and worst was assigned a 0

Discussion

First, consistent with Kornell et al.'s (2009) study, we showed that producing an error for semantically related materials led to enhanced retention. We also found that error-generation did not enhance recall if the materials were completely unrelated. The semantic relation between the cue and target appeared to be critical in determining whether error-generation enhanced memory or not.

A question one might ask is whether participants were behaving similarly when they generated their errors and responded to the feedback in the related and unrelated materials conditions. Perhaps participants were simply guessing randomly and were not sufficiently engaged in the unrelated materials condition, while they were employing all of their efforts to try to generate the answers in the related materials condition. An attentional explanation has been proposed in other error correction paradigms (Butterfield & Mangels, 2003 Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009). Izawa (1967, 1970) has specifically argued that previous errors led to increased learning because of enhanced attention to the corrective feedback. Motivational/attentional differences between conditions might be revealed by the nature of their guesses. By examining the nature of the error responses that the participants produced, we could potentially gain some insight into whether participants' behavior was substantively different behavior was when they generated their errors in the related and unrelated materials conditions.

Latent semantic analysis We obtained estimates of the relation between the cues and the generated errors by using latent semantic analysis (LSA). LSA (see Landauer, Foltz & Laham, 1998) is a method of extracting the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). The aggregate appearance of all words provides a set of mutual constraints that is thought to determine the similarity of meaning of words to one another, given as a cosine. Using LSA (through http://cwl-projects.cogsci.rpi.edu/msr/; see Veksler, Grintsvayg, Lindsey & Gray 2007), it was found, as expected, that the mean relatedness between the cues and targets was higher for related materials (M = .27, SD = .04), than for unrelated materials (M = .05, SD = .01), $t(58)$ = 30.46, SE = .01, p < .001. Of more interest, we used LSA to investigate the association between the cue and the error that was generated, in the related and unrelated materials conditions. As is shown in Table 2, when presented with the cue, participants produced errors that were related to the cue in both the related and the unrelated materials conditions. The mean relatedness between cue and generated error for related materials (M = .28, SD = .09) was numerically only slightly higher than that for unrelated materials (M = .25, SD = .09), $t(58)$ = 1.92, SE = .02, p = .056). To see whether participants in the unrelated materials condition altered their guessing strategy as the experiment progressed, the mean association values for the first 15 items were compared with those for the last 15 items. There was no difference in the LSA values for the later trials (M = .24, SD = .09) from those for the earlier trials (M = .26, SD = .08), t = 1.09.

Table 2 Latent Semantic Analysis (LSA, a semantic relation tool) enabled analysis of the semantic relatedness between the Errors produced by the participant to the provided Cues and Targets. Mean cosine values (the measure provided by LSA) between word pair comparisons are presented below. Higher values indicate a higher degree of semantic relation. These data are presented as a function of Materials condition and Accuracy on the final cued recall test. Standard deviations are provided in the parentheses

	Cue to Error		Target to Error	
Materials	Correct	Incorrect	Correct	Incorrect
Experiment 1				
Related	0.30 (.07)	0.25 (.10)	0.20 (.04)	0.19 (.10)
Unrelated	0.25 (.11)	0.25 (.08)	0.09 (.04)	0.07 (.02)
Experiment 2				
Related	0.27 (.09)	0.27 (.07)	0.20 (.13)	0.22 (.06)
Unrelated	0.34 (.17)	0.32 (.06)	0.06 (.05)	0.06 (.02)

As was noted in the introduction, we hypothesized that the relation between the generated error and the target might be a critical factor in determining whether error-generation would be beneficial for memory—a possibility that we could also investigate using LSA. Table 2 shows the mean association values for the target–error relation as a function of materials. And indeed, as was hypothesized, the error was more related to the target in the related materials condition (M = .20, SD = .04) than in the unrelated materials condition (M = .08, SD = .02), $t(58)$ = 17.16, SE = .01, p < .001.

Metacognitive illusion The metacognitive results were particularly interesting. These retrospective judgments were taken after the participants had already had considerable experience with the task. Although participants had just completed the final test moments earlier, those participants in the related materials condition did not realize that the error-generation condition led to the best performance. Instead, they erroneously thought that the read-long condition was the most beneficial for memory of the target items—failing, rather dramatically, to appreciate the benefits of making errors. Furthermore, although performance in each of the three different learning conditions varied greatly between materials, the metacognitive ratings were similar. Comparing materials conditions, what is clear is that although the performance follows two distinct patterns, the metacognitive ratings do not vary as a function of material relatedness. The metacognitive rankings for each learning condition (read-short, read-long, and error-generation) revealed no statistical differences across materials, (zs < 1.60, ps > .13). Therefore, although we see a performance boost from error-generation for related materials, participants' rankings are no different from those in the unrelated condition. This metacognitive illusion, it seems, is stable and unaffected by the participant's own contradictory experience with the results of the learning task.

Experiment 2

In the second experiment, we endeavored to replicate the results of Experiment 1 in a within-participants design, to address more fully the question of why there was a benefit of error-generation only when the cue and target were semantically related. One motivation for a within-participants design was that randomly mixing the presentation of related and unrelated materials would ensure that participants were cognitively engaging in similar tasks when generating an error and would obviate the small difference in response to the cues seen in the LSA analysis in Experiment 1. In the within-participants design, when only the cue was displayed on the screen, the participants could not know whether the forthcoming target would be related or unrelated to the cue. If the lack of memorial benefit for unrelated materials from error-generation was an artifact only of overall lack of engagement or attention, a benefit of generating errors might occur for both related and unrelated materials in the within-participants design. Only after error-generation could participants know the relation of the cue and the target. Conversely, if we replicated the results seen in Experiment 1, this would provide stronger evidence that the semantic relation between the error and the target is central in determining when error-generation helps memory.

Method

Participants Thirty-six Columbia students participated for credit. Six nonnative English speakers were excluded, leaving 30 participants' data. Mean age was 20.7 years (SD = 3.3), and 50% of the participants were females.

Design and materials A 2(materials: related, unrelated) × 3 (learning: read-short, read-long, error-generation) within-participants design was used. Forty-five of the related material items and 45 of the unrelated material items from Experiment 1 were randomly selected for use in the present experiment, for a total of 90 word pairs. For both related and unrelated materials, three sets of 15 word pairs were created and counterbalanced over participants so that each word pair was assigned to each of the three learning conditions equally.

Procedure The procedure was the same as that in the previous experiment. During the study phase, item presentation order was randomized, and, as noted above, items were preassigned to conditions, which were counterbalanced between participants. Order of item presentation was also randomized on the final cued recall test. All instructions were identical to those given in Experiment 1, with the exception of the metacognitive ratings. Since the present design had six conditions, all six were described in the instructions before the participants ranked them in order of best final test performance to worst.

Results

Learning phase performance Participants did not correctly answer any of the unrelated materials in the error-generation condition during the learning phase. They correctly guessed 3% of the related targets (SD = .03). All results from the error-generation condition excluded the trials for which participants guessed correctly on the initial test.

Final cued recall test performance As is shown in Fig. 1, there was an interaction between learning condition and materials. Error-generation led to the highest correct performance for related materials, but it did not lead to benefits with unrelated materials, $F(2, 58) = 7.89$, MSE = .01, $p < .01$, $\eta_p^2 = .92$. Pairwise comparisons showed that error-generation for related materials led to higher correct recall (M = .70, SD = .20) than did both read-short, $t(29) = 4.08$, SE = .04, $p < .001$, and read-long, $t(29) = 2.51$, SE = .04, $p < .05$, which did not differ from one another, $t(29) = 1.61$, SE = .04, $p = .12$. There were no significant pairwise differences in performance for the three learning conditions with unrelated materials (all ts < 1). As was expected, participants remembered more of the correct targets for the related than for the unrelated materials, $F(1, 29) = 344.32$, MSE = .03, $p < .001$, $\eta_p^2 = .21$. Although qualified by the interaction, there was a main effect of learning condition such that error-generation led to the highest correct performance overall, followed by read-long and read-short, $F(2, 116) = 4.06$, MSE = .03, $p < .05$, $\eta_p^2 = .12$.

Reaction times Table 1 shows mean RTs as a function of accuracy on the final cued recall test, learning and material conditions. Only 16 participants had observations for all cells. Overall, items answered correctly (M = 3.69 s, SD =

1.34) were produced more quickly than incorrect items (M = 7.04 s, SD = 2.84), F(1, 29) = 69.18, MSE = 14.59, p < .001, η_p^2 = .71. When participants previously made an incorrect guess in the error-generation condition (M = 6.01 s, SD = 2.25), their subsequent RTs on the final cued recall test were longer than those in the read-short (M = 5.02 s, SD = 2.19) and the read-long (M = 5.08 s, SD = 1.82) conditions, F(2, 58) = 6.88, MSE = 5.31, p < .01, η_p^2 = .19. Items in the error-generation condition that were answered incorrectly on the final test took longer to produce than items answered correctly, F(2, 58) = 5.73, MSE = 4.35, p < .01, η_p^2 = .17. The relatedness of the materials did not lead to differing response latencies on the final test, F < 1, nor did materials interact with any other factor.

Error persistence For unrelated materials, 15% of the responses on the cued recall test were original errors that persisted from the learning phase to the final test. The original errors persisted only 7% of the time for related materials, t(29) = 3.65, MSE = .01, p < .01.

Metacognition Performance for each condition was ranked from best to worst. Because there were six conditions in the experiment, the best condition for each participant was assigned a score of 5, and the worst was assigned 0. As can be seen in Fig. 3, the error-generation condition for related materials was objectively the best condition for retention (M = 4.38, SD = .87). However, this condition was given only a mean metacognitive ranking of 2.53 (SD = 1.54), z = 4.12, p < .001. Conversely, although read-long for related materials was subjectively believed to have produced the best performance (M = 4.53, SD = 1.03), in fact, it most often led to worse performance than did error-generation (M = 3.72, SD = 1.07). Noticeably, the metacognitive ranking and performance ranking for read-long are not aligned, z = 3.34, p < .001. Finally, mean metacognitive judgments indicated that participants subjectively believed that the unrelated read-long condition led to better performance than it actually did ($M_{\text{metacognitive}}$ = 2.37 SD = 0.82; $M_{\text{performance}}$ = 1.37 SD = 0.97), z = 3.48, p < .01. No significant differences were found between the mean metacognitive and performance rankings for the three other cells (related–read-short, unrelated–read-short, and unrelated–error-generation), zs < 1.

Latent semantic analysis The results from the LSA mirror those of Experiment 1, despite many participants being excluded from the analysis due to lack of observations in every cell (see Table 2). Participants generated errors that were related to the cue regardless of materials condition. The mean relation value between the cue and error was slightly higher for unrelated materials (M = .32, SD = .06) than for related materials, (M = .26, SD = .05) t(29) = 5.06, SE = .01, p < 001, although at the time of generating the error, the participant could not be aware of the subsequent relation to the target, and this effect is in the opposite direction in Experiment 1. The semantic relatedness of the errors to the cues provided support for the idea that participants were engaged and truly generating reasonable errors, even for the unrelated materials. As was expected, errors in the unrelated materials condition were not as related to the target (M = .06, SD = .02) as were errors generated for related materials (M = .21, SD = .06), t(29) = 17.51, SE = .01, p < .001.

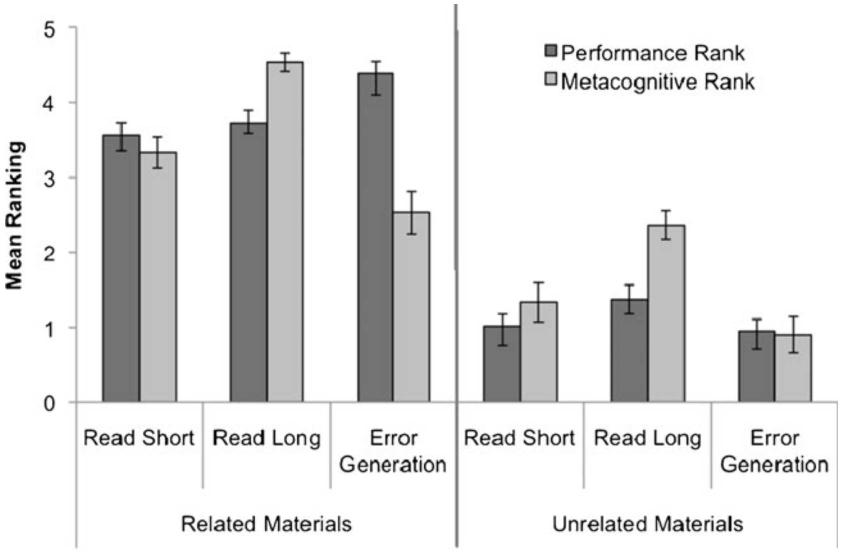

Fig. 3 Metacognitive Data for Experiment 2 (within-participants). Mean ranking of Learning conditions based on correct performance on the final cued recall test and subjective metacognitive judgments. The condition with the highest proportion correct or subjectively rated the best was assigned a score of a 5. Second best was assigned a score of 4 and so forth, while the worst condition was assigned a 0

Discussion

Experiment 2 replicated the findings of Experiment 1: Error-generation led to memorial benefits over both reading conditions, but only for related materials. For semantically related word pairs in both experiments, there was enhanced retention for the correct response when participants had made a prior incorrect response, as compared with when they had just read the word pairs. For the unrelated materials, there was no such benefit of incorrect guessing in either experiment. If the benefit from producing an error was due to the effort or engagement during generation itself, there should have been some benefit from incorrect guessing for the unrelated materials. Insofar as items were randomized, in the second experiment, participants could not have been aware of what the next trial would be. Therefore, the processing was the same across related and unrelated conditions during the act of generating the error itself. The results from the LSA substantiated this lack of difference during error-generation. Therefore, it appears that the differential benefits of error-generation between related and unrelated materials began at the time of target feedback.

General discussion

These results support the idea that semantic closeness is a critical factor in determining whether an error will or will not help learning. One framework consistent with these results is the Osgood (1949) transfer surface, which captured all transfer of learning relations that were known at the time of publication. In this surface, similarity between intralist stimuli (cues) is plotted against similarity of intralist responses (targets). Of importance is how these two factors interact to produce positive transfer or, conversely, interference. When cues are identical, as in our experiments, the more related the responses are to one another, the greater the positive transfer will be. Thus, this framework would predict that positive transfer will result if the error and target are related. It is only when the two responses—the error and the target—are unrelated that negative transfer or interference should be produced. For the error-generation condition for related materials in our experiments, LSA showed that errors were highly similar to the correct targets. Therefore, Kornell et al.'s (2009) materials and our related materials condition conformed to Osgood's (1949) A–B A–B' situation. The erroneous answer produced in this context facilitated learning of the correct answer, B'. Conversely, the LSA ratings showed that our unrelated materials condition conformed to Osgood's A–B, A–C situation. The errors, in that condition, were unrelated to the correct targets and produced no memory benefit. In fact, there was a slight suggestion of error-related interference. As compared with related materials, unrelated materials led to more of the original errors persisting in the final test. Additionally, in the first experiment, correct item recall was worse in the error-generation condition than in either the read-long or read-short condition.

Since the time of Osgood (1949), two possible explanations have emerged for why this relationship between the error and the target might be important. One explanation is that making a related error helps form a richer, more elaborate network with the cue and the error, as compared with an unrelated error. In terms of levels of processing, encoding in a deeper, more elaborative manner is beneficial for future retrieval (Craik & Lockhart, 1972; Craik & Tulving, 1975). Through elaborative processing, by producing a guess and forming an elaboration based on a "deep" or semantic level, retention is enhanced above "shallow" processing. Error-generation of a related item might be an elaboration, making the target more meaningful. Although one might engage in elaborative processes for unrelated materials, this elaboration might be in vain. For example, when provided with the word *attack*, when one tries to generate a response, one will presumably think about what it means, and generate, erroneously, *dog*. When the related target, *defend*, is displayed, the connection is clear, and one can draw a more elaborate and meaningful relationship than when one simply sees *attack–defend*. One can imagine an attack dog defending his doghouse, defending oneself against an attacking dog, or both. This richer, more elaborate encoding method should help retention. However, if the correct answer is something unrelated to attack, such as *bicycle*, it is more difficult to form a meaningful connection or elaboration between the cue, error, and target. Additionally, Carpenter (2009) and Carpenter and DeLosh (2006) have argued that elaboration is less likely to occur when one is reading, as compared with active retrieval.

Along similar lines, during error-generation, one might activate a variety of related concepts that provide a more elaborate, richer memory trace, consistent with spreading activation theories of memory (e.g., Collins & Quillian, 1972). Since there is more information that could potentially activate the correct target, this elaborative structure could aid recall (e.g., J. R. Anderson, 1983). Carpenter (2011) suggests that retrieval aids in activating semantically related information above restudy. In other recent work, Grimadli and Karpicke (in press) found error-generation benefits only for semantically related items, a finding consistent with the results presented in the present article. Conversely, when participants' errors were constricted by providing the first few letters of the error (e.g., *tide–wa*____), an error-generation benefit was not obtained. The authors interpret their results as favoring a spreading activation view—that is, when an error is committed, concepts that are related to the target are activated and enhance learning (e.g., Collins & Loftus, 1975).

A mediator hypothesis is a second potential explanation for why the relation between the error and the target may be important in determining whether errors benefit recall of the correct target. Under some circumstances, the error itself may serve as a mediator, or secondary link, between the cue and the target. It has been shown that previous retrieval attempts can serve as an intermediary cue in target retrieval (Soraci et al., 1999) and can facilitate recall (Pyc & Rawson, 2010). The latter authors found beneficial effects of the mediator, however, only when it both could be retrieved at time of test and elicited the target item.[3] In the present paradigm, it seems more likely that a word that is related to the target might serve as an effective mediator than would one that is unrelated to the target.

These two hypotheses—*error as an elaboration* and *error as a mediator*—are not mutually exclusive. *Error as an elaboration* suggests that because of enhanced processing at encoding from an active (elaborative processing) or passive (semantic activation) process, the correct target will be remembered better when an error is generated than with simple study. In addition, even at retrieval, those concepts that were previously activated might lead to enhanced recall of the correct target. On the other hand, the *error as a mediator* hypothesis suggests that recalling the original error itself, and not just the surrounding semantic landscape, can act as a secondary cue to retrieve the target.

The RT data are readily interpretable within the *error as a mediator* hypothesis. Participants took longer to produce a response on the final test for the error-generation condition, as compared with the read-long and read-short conditions. When they attempted to retrieve the correct target, the incorrect guesses might have served as a secondary link that introduced a second step into the retrieval process. This second step would require additional time, thereby leading to longer RTs. Even for unrelated materials, if one retrieved the original error and tried to use it as a mediator, the response time would still be longer due to the additional, unsuccessful labor involved in trying to find the correct retrieval path to the target.

The RT data could also be interpreted within the *error as an elaboration* view, though, insofar as exploring the elaborations that were set up at encoding could be assumed to take time. A number of semantic activation models predict longer RTs with a higher number of associated concepts (see ACT–R and the fan effect; J. R. Anderson, 1974; J. R. Anderson & Reder, 1999). These models could also predict that participants' RTs would be longer for the error-generation conditions as a result of response competition between the original generated error and the correct target.

Finally, in both experiments, the metacognitive data show a stable illusion, whereby participants were not aware that error-generation was helpful for remembering related word pairs. It is, perhaps, not surprising that committing errors during learning is typically seen in a negative light. As Bjork (1994) stated, "Errors made during training are generally not viewed as opportunities for learning, but rather, as evidence of a less-than-optimal training program" (p. 299). It is surprising, however, that even moments after completion of the criterion test, participants were not aware that error-generation was beneficial for related materials. This finding is particularly interesting since these global retrospective judgments of performance can use information acquired during the criterion test to help inform judgments. Retrospective judgments, therefore, have been shown to be more accurate than predictions of performance (see Pieschl, 2009). For this reason, it is particularly interesting that there seems to be such a large disconnect between subjective performance rankings and actual performance.[4]

Although, currently, we cannot make any claims in regard to potential mechanisms driving the subjective bias against errors, this bias is still of great interest. There are several possible explanations for the error-generation metacognitive illusion. One is that participants simply had a bias against believing that errors are beneficial. A second explanation is that participants relied on a *ease of processing* heuristic (see Koriat & Ma'ayan, 2005; Winkielman, Schwarz, Fazendeiro & Reber, 2003) or, more specifically, *easily learned, easily remembered* (Koriat, 2008; Miele & Molden, 2010). There have been a number of experiments in which how easily stimuli are processed influences judgments of how well information is learned (e.g., Carpenter & Olson in press; Koriat, 1997, 2008; Nelson & Dunlosky, 1991; Rawson & Dunlosky, 2002; Rhodes & Castel, 2008). Since error-generation might not have seemed as easy (perhaps both at retrieval and at encoding) as reading the answer, participants might be underconfident in this strategy. Furthermore, if participants were also generating the error as a mediator, despite its beneficial effect on retention, the presence of another potential competitor could have driven down performance estimates.

From an educational standpoint, the findings of the two reported experiments are of relevance for two reasons. First, we have shown that when the materials are related, even when that relation is very small—low associates, not high associates—

[3] Although more original errors were produced on the final test for unrelated materials in Experiment 1, this does not necessarily mean that those in the related materials condition were not capable of retrieving their original error. Anecdotally, during the debriefing, many participants in the related materials condition mentioned that they remembered their guesses.

[4] It is possible, from the present data, that participants underestimated the memorial benefits of generating errors because they could not remember which items were in the error-generation condition (although cf. Huelser & Metcalfe, 2011).

generating an error and receiving corrective feedback is much better for learning than is simply studying. Although more research must be done to understand the exact mechanisms behind the error-generation effect, the present results suggest that guessing should be encouraged, even if the result is an error. Rarely will the question and answer be so far removed that the learner cannot make a meaningful connection between the two. However, some caution is needed in implementing this recommendation, given that errors may have detrimental effects for memory-impaired individuals, as was shown in Clare and Jones's (2008) review. It is not yet known whether error-generation, when the errors are related to the targets, as in the present study, will lead to enhanced or diminished performance for young children or those with learning disabilities.

The second point of interest to educators comes from the metacognitive monitoring results. It is clear that even immediately after completion of the criterion test, participants were not aware of which study strategy was best for learning. It is quite plausible that learners rely on these types of global retrospective judgments when deciding what learning strategy to use. It has been shown that monitoring has consequences for metacognitive control, or regulation, of learning (Metcalfe & Finn, 2008; Son, 2004; Son & Kornell, 2008; Son & Metcalfe, 2000; Stone, 2000; see also Metcalfe, 2009). Thus, it seems unlikely that the learner, without further training of his or her metacognition, will implement this highly effective learning strategy.

Author note This research was supported by Grant 220020166 from the James S. McDonnell Foundation, and by a National Science Foundation graduate fellowship to the first author. We thank Brandon Luke and Margaret Lee for assistance with material creation and data collection. We also thank Lisa Son, Karen Kelly, David Miele, Peter Messa, Joseph Bisoglio, Katherine Rawson, Jeff Karpicke, Sean Kang, and an anonymous reviewer for their helpful comments.

References

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review, 80,* 417–438.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6,* 451–474.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Behavior, 22,* 261–295.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123.

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory: Handbook of perception and cognition* (2nd ed., pp. 237–313). San Diego, CA: Academic Press.

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

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

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

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A metanalytic review. *Memory & Cognition, 35,* 201–210.

Bjork, R. A. (1994). Institutional impediments to effective training. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 295–306). Washington, DC: National Academy Press.

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

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

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

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547–1552

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276.

Carpenter, S. K., & Olson, K. M. (in press). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review, 1,* 1–23.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Collins, A. M., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L. Gregg (Ed.), *Cognition and learning* (pp. 117–138)). New York: Wiley.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 10,* 268–294.

Cunningham, D., & Anderson, R. C. (1968). Effects of practice time within prompting and confirmation presentation procedures on paired associate learning. *Journal of Verbal Learning and Verbal Behavior, 7,* 613–616.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89,* 627–661.

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

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Glaser, R. (1990). The reemergence of learning theory within instructional research. *American Psychologist, 45,* 29–39.

Grimadli, P. J. & Karpicke, J. D. (in press). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*.

Guthrie, E. (1952). *The psychology of learning* (Revth ed.). New York: Harper.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16,* 96–101.

Huelser, B. J. & Metcalfe, J. (2011, November). Performance monitoring offsets, but does not eliminate, the metacognitive illusion that errors hurt learning. *Poster presented at the 52nd annual meeting of the Psychonomic Society, Seattle, WA*.

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology, 75,* 194–209.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344.

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

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70,* 626–635.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology., 131,* 48–59.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition, 36,* 416–428.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52,* 478–492.

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.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414.

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

Melton, A. W., & Irwin, J. M. (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.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145–160.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18,* 159–163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15,* 174–179.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review, 14,* 225–229.

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

Miele, D. B., & Molden, D. C. (2010). Naïve theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General, 139,* 535–557.

Mulligan, N. W., & Lozito, J. P. (2005). Self-generation and memory. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 175–214). San Diego: Elsevier.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL-effect." *Psychological Science, 5*, 207–213.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from http://w3.usf.edu/FreeAssociation/

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56,* 132–143.

Parlow, J., & Berlyne, D. E. (1971). The effect of prior guessing on incidental learning of verbal associations. *Journal of Structural Learning, 2,* 55–65.

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.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1051–1057.

Pavio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76,* 1–25.

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition Learning, 4,* 3–31.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28,* 69–80.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137,* 615–625.

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.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604.

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

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 601–604.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 204–221.

Soraci, S. A., Jr., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory & Language, 41,* 541–559.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12,* 437–475.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73.

Veksler, V. D., Grintsvayg, A., Lindsey, R., & Gray, W. D. (2007). A proxy for all your semantic needs. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (p. 1878). Austin, TX: Cognitive Science Society. Retrieved from http://cwl-projects.cogsci.rpi.edu/msr/

Webb, L. W. (1917). Transfer of training and retroaction: A comparative study. *Psychological Monographs, 24,* 1–90.

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Erlbaum.