Learning by making errors: When and why errors help memory, and the metacognitive illusion that errors are hurtful for learning

Barbie Jean Huelser

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Learning by making errors: When and why errors help memory, and the metacognitive illusion that errors are hurtful for learning

Barbie J. Huelser

This body of work begins to investigate the following three overarching questions on errors and learning. First, when are errors helpful for memory? Second, why are errors beneficial in certain circumstances? Third, are learners aware of when errors are advantageous for learning?

These questions cover two unique dimensions of learning by making errors, both from a memory and a metacognitive point of view. From a memory perspective, it might seem surprising that making an error compared to simply studying (no mistakes) could be beneficial for memory. We began our investigations with a replication and extension of previous work on the *error generation effect*: When does making errors enhance correct retention above studying? By investigating boundary conditions, this helps inform theory of the mechanism responsible for the error generation effect. We found that error generation only enhanced retention for related materials but not for unrelated word-pairs, and therefore, confirm that the error generation benefit is more than simply due to the act of generation.

However, what is the role of the error: Does it serve as a semantic mediator linking directly to the semantically related target, or can the error serve as an episodic link, bridging to the original learning episode, even if it is not directly linked to the target? If a learner remembers her error, does this help or hurt memory for the correct answer? By using materials that enabled errors that were either congruent (related) or incongruent (unrelated) to the correct answer, we

found generating errors during learning led to benefits of memory, both when the error was congruent and incongruent to the target. Furthermore, when one could recall her error at test, correct answer memory was higher than when one could not recall her original error. These findings suggest that just a semantic explanation for the error generation is likely insufficient, and point to the importance of episodic recollection at retrieval for error generation to aid memory above study alone.

Lastly, we investigated this errorful learning methodology from a metacognitive perspective. Even when errors were beneficial for learning, we found that learners were unaware of the memorial advantage. We sought to ensure this underconfidence was not merely a function of poor performance accuracy or source monitoring. We were also interested in exploring if this bias was stable, or if one could correctly update her metacognitive knowledge simply by making item-level judgments. These initial projects open the doors for exciting research investigating individual differences on learning from generating errors, from both a memory and metacognitive perspective.

Table of Contents

Lists of Charts and Graphs

Acknowledgments

This research was supported, in part, by Grant 220020166 from the James S. McDonnell Foundation and the National Science Foundation Graduate Student Research Fellowship. I would like to thank all the past research assistants of the Metcalfe Lab, particularly Kelsey McLeod, Michael Smith, Keren Fefer, Margaret Lee, Rachel Burris, Christina Crowther, Erica Tanne, Lyla Parvez, Anna Fischel, Mark Rhodes, Brandon Luke, Corey Fernandez, Erica Baruch and Zachary Bucknoff. I would also like to thank past and present lab members, Dr. David Miele, Dr. Lisa Son, Judy Xu, Matti Voure, and especially Dr. Karen Kelly. I would also like to acknowledge the support of my committee and extend my gratitude for their valued advice and time: Drs. Tory Higgins, Kevin Ochsner, Daphna Shohamy and Katherine LaTour. Most of all, I would like to thank my advisor and mentor, Dr. Janet Metcalfe. Her contributions and support have been immeasurable, and this work would not be possible without her unending guidance and encouragement.

Dedication

I dedicate this culminating work to Women in Science, especially my first memory professor, Dr. Rebekah Smith. She was the first to spark my interest in the field of memory and learning from errors, thanks to her challenging pre-lecture assignments that I often answered incorrectly. On a personal note, I dedicate this work to my mother, Janice Cody, my father, Steven Huelser, my sisters, Jill and Britnee Huelser, and to my husband and best friend (and spell-checker), Peter Messa. Without their unwavering and unconditional support, this endeavor would not have been possible.

Preface

Recent research has shown that producing an error, so long as it is followed by corrective feedback, resulted in better retention of the correct answers than simply studying the correct answers. We wanted to investigate this surprising finding from two vantage points: memory and metacognition.

In the first two chapters, we consider this puzzle of error generation from a memory mechanism point of view. In Chapter 1, we explored a replication of this finding in which the errors were always related to the target, and examined whether there are situations in which error generation might not be beneficial, in particular, when the errors were unrelated to the to-be-remembered target item. By investigating boundary conditions for this effect we could begin to tease apart if errors were helpful merely due to active processing or because of the attention paid to the feedback post-error. In either case, one would expect the error generation effect for both related and unrelated materials. However, if errors are important either due to mediation or semantic activation of the correct response, we should expect an error generation effect only for related materials. In Experiments 1a and 1b, participants studied either related (Experiment 1a) or unrelated word pairs (Experiment 1b), manipulated between participants. Participants were either given the cue and target to study for 5 s or 10 s in the read conditions, or they generated an error in response to the cue for the first 5 s before receiving the correct answer for the final 5s. 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. Experiment 1c replicated these findings in a within-participants design. As will be discussed in greater detail in Chapter 3, learners did not

know that generating an error enhanced memory, even after they had just completed the task that produced substantial benefits.

In Chapter 2, we sought to elaborate further upon possible mechanisms of the error generation effect. One explanation is based on how closely related the error is to the correct answer (semantic mediation hypothesis). We tested the role of the semantic link between the error and the target item by using polysemous materials to create congruent (wrist-palm-hand) and incongruent (tree-palm-hand) triplets. Participants generated errors that were congruent/related (cue: wrist-palm-__?__; error: finger) or incongruent/unrelated (cue: tree-palm-__?__ ; error: coconut) to the correct answer (hand). A benefit for error generation was found in the congruent condition, as expected. However, even in the incongruent condition, when there was no semantic link between the mistake (coconut) and correct answer (hand), error generation was still beneficial for correct target memory (Experiment 2a). This advantage only occurred, however, when the original error was also recalled on the final memory test (Experiment 2b). These novel findings do not support the Semantic Mediation account as the sole explanation for the error generation effect. Instead, we propose the Episodic Recollection hypothesis: making an error can serve as an episodic memory link to the correct answer.

Chapter 3 addresses the new finding from Chapter 1, that learners are unaware that error generation is beneficial for memory under many circumstances, and attempts to answer whether or not these impressions are fixed or malleable. Our aim was to try to enable learners to overcome this metacognitive disconnect. We used similar methods and materials as in Chapter 1 (Experiment 1a). However, in Experiment 3a, a between participants manipulation was introduced on the final cued recall test. The control (no monitoring) group completed the final test without an additional task, which was akin to the procedure in Chapter 1. The Experimental

group (confidence monitoring) monitored performance on an item-level by providing a confidence rating of accuracy for each response on the final test. After completing the cued recall test, all participants made a subjective Global Retrospective Estimate of Performance (GREP) for each learning condition (read short, read long, error generation). Though error generation produced the best correct retention, both groups were underconfident in their retrospective performance estimates. The error generation metacognitive illusion was reduced, though not eliminated, through the use of item-level performance monitoring (in the Experimental group of confidence monitoring). In Experiment 3b, in addition to confidence judgments, half of the participants were also asked on the recall test to indicate how each item had originally been presented during learning. The design was a (2 (confidence monitoring on cued recall: yes, no) x 2 (source monitoring on cued recall: yes, no) X 3 (learning condition: read short, read long, error generation) [within-participants], mixed design). Monitoring confidence or source during test led to greater global retrospective estimates of performance (GREPs) than those who did not monitor their performance or how an item was originally studied. Monitoring also had consequences for future learning strategy selection; Monitoring during the criterion test led to a greater number of error generation items selected for a future test, compared to when no overt monitoring occurred during recall. In summary, performance and source monitoring at the item-level aided in updating metacognitive knowledge about the effectiveness of learning by making errors.

This collection of work covers only a small portion of possibilities that will elucidate when, why and for whom, generating errors can be more effective for memory than just studying.

Chapter 1

The Impact of Errors on Memory

The Impact of Errors on Memory

This first chapter addresses the effect of making errors on learning. Should one learn by studying materials without making mistakes, or by attempting to produce the answers and committing the 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 & Roher, 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., 2003; Pashler et al., 2005). In this chapter, 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 chapter is the memorial consequences for errorful as compared to errorless learning, a related question of interest is: 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 (Metcalfe & Finn, 2008; Kornell & Son, 2009). If one is not aware of the potential efficacy of a learning strategy, the learner might implement suboptimal strategies. Hence, one'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 the cue A is later paired with a different response (C) (Anderson & Neely, 1996; Anderson & Reder, 1999; Barnes & Underwood, 1959; Loftus,

1979; Melton & Irwin, 1940; McGeoch, 1952; Osgood, 1949; Webb, 1917). Though there are several theories concerning how this interference arises (e.g., Anderson, 1973; 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 during learning be eliminated (Glaser, 1990). For example, Guthrie (1952) suggested that errors should be avoided because by practicing errors, the incorrect response to a particular stimulus would be strengthened. Furthermore, errorless as compared to 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 specific patient populations and may not apply to typical learners. Nevertheless, in an experiment by Cunningham and Anderson (1968), worse retention was found after 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 to allowing participants 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, and also see Kornell & Metcalfe, 2014) found neither benefit nor impairment for forced as compared to free responding. Similarly, Kang et al. (2011) found that forced guessing did not lead to either better or worse memory for the correct answer on a later retention test, either immediately or at a one-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, Roediger & McDermott, 2011). No difference in memory was found for items quizzed on a pretest as compared to non-quizzed items.

In contrast to the above findings, however, there are some studies in which making errors helps learning. Richland, Kao and Kornell (2009) found enhanced memory for material from reading passages when the to-be-remembered material was tested using cued recall questions prior to reading 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 to 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 either in a Reading mode or 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 5s or 13s). In the Error-generating mode, participants only saw 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 to 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, which demonstrated 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 to 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 will be 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 if, in retrospect, participants were able to accurately monitor whether generating errors helped or hurt their performance on the final test. This question is important, as metacognitive monitoring has been shown to have consequences for strategy selection, referred to as metacognitive control (Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003).

Experiments 1a and 1b

In Experiment 1a, our aim was to replicate Kornell et al.'s (2009) findings by investigating if we would also find an error generation effect for memory of weakly associated word pairs. Participants studied word pairs in an error generation condition, and two different read conditions (within participants factor). In Experiment 1b, we extend upon these findings by utilizing unrelated word pairs, for which we hypothesized the effect would not be found. We present the data for these two experiments as a between participants factor.[1] Therefore, 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.

Methods

Participants. Sixty Columbia undergraduates (native English speakers) participated for partial fulfillment of a class requirement. 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) x 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's (1998) norms, closely following Kornell et al.'s (2009) word

[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.

pair selection criteria. Given the first word, approximately 5% of participants in Nelson et al.'s (1998) experiment produced the target as the first associate. Specifically, Forward Associative Strength was between .05-.054, and Backward Associative Strength was 0. Each word was a minimum of 4 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's (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 4 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. Mean concreteness ratings were the same for the words assigned as cues and targets (M = 6.77). For each of the between-participant 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 only given 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 seconds, 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 10s. These conditions were presented in a random order (not blocked). The computer made a soft clicking sound to alert the participant 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 minutes 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 based on 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 5s, B) together – long: both words displayed on the screen for 10s; C) separate: the first word presented separately (5s) before both words were displayed (5s). 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 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 test 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, therefore, 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 Figure 1.1, correct final performance was higher for related materials (M = .64, SD = .19) compared to 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 lead 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 interaction with type of Materials. Although error generation enhanced

[2] Of these errors, 90% were errors of commission for related materials, and 91% for unrelated materials, $t < 1$. Reported data is including 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 1c, 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 only be computed for generated errors).

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 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 both the read short condition ($M = .21$, $SD = .16$), [$t(29)$ 2.09, $SE = .02$, $p < .05$], and the error generation condition ($M = .17$, $SD = .12$), [$t(29) = 3.26$, $SE = .02$, $p < .01$]. Though the trend favored the read short condition over error generation, performance between these two conditions was not significantly different from one another, [$t(29) = 1.89$, $SE = .02$, $p = .068$].

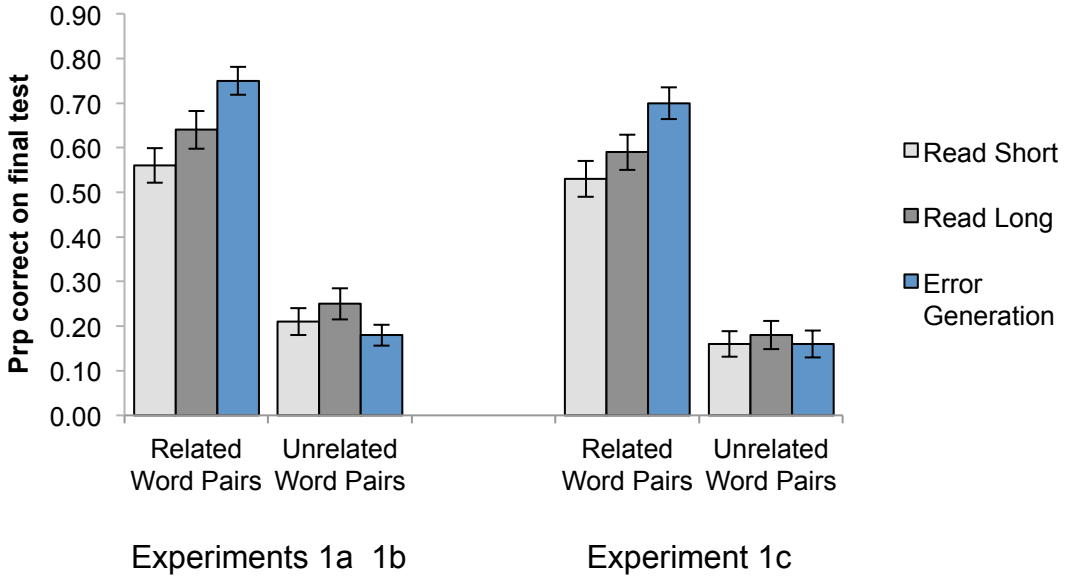


Figure 1.1. **Cued Recall Performance.** Correct performance on final cued recall test as a function of Learning condition and Materials for both Experiment 1ab (between-subjects) and Experiment

Reaction times. Reaction time (RT) data on the final test were analyzed as a function of Accuracy on the final test (correct versus incorrect), learning condition (read short, read long, error generation) and materials (related, unrelated); see Table 1.1 for means. Reaction time data are reported in the current section for completeness, but will be discussed only in the General discussion. 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.

	Correct on Final Test			Incorrect on Final Test		
	Read Short	Read Long	Error Generate	Read Short	Read Long	Error Generate
Experiments 1a 1b						
Related (1a)	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 (1b)	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 1c						
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)

Table 1.1
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 the parentheses.

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 to 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 generate 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 to related materials ($M = .05$, $SD = .06$), [$t(58) = 4.05$, $SE = .04$, $p < .001$].[3]

Metacognition. Data from 52 subjects 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 on which the participant performed best on the final test was assigned a 2; the condition on which he or she performed second best was assigned a 1; and the worst was given a score of 0. The same assignment was done for individuals' metacognitive ratings of the three Learning conditions.

[3] Though more of original errors were produced on the final test for unrelated materials in Experiment 1a, 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.

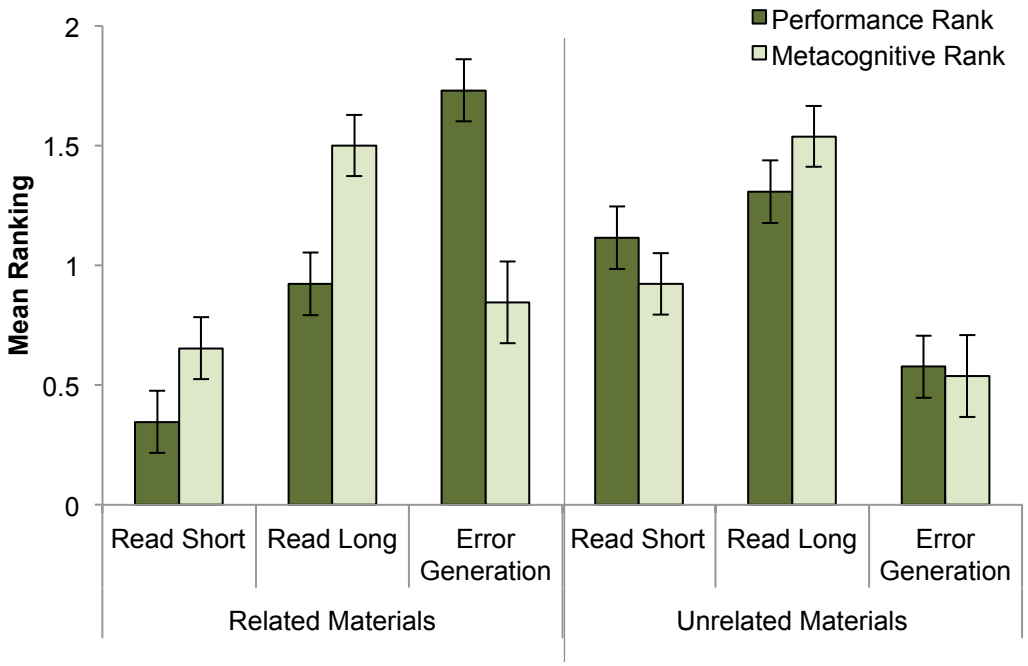


Figure 1.2. Metacognitive Data for Experiments 1a1b (between-participants). Mean ranking of Learning conditions based on correct performance on the final cued recall test and subjective metacognitive judgments. The condition with the highest proportion correct or subjectively rated the best was assigned a score of a 2. Second best was assigned a score of 1, and worst was assigned a 0.

As can be seen from Figure 1.2, 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 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 non-parametric 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 = .35, SD = .56) to be worse than subjectively reported (M = .65, SD = .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 = .92, SD = .61), [z = 2.78, p < .01]. Most interestingly, however, in the error generate condition participants mistakenly believed that their performance was very low (M = .85 SD = .88) when it was actually high (M = 1.73, SD = .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 compared to actual performance rank for any of the comparisons, [zs < 1.16].

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 (c.f., Butterfield & Mangels, 2006, Butterfield &

Metcalfe 2006; Fazio & Marsh, 2008). 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 had behaved substantively differently 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 of appearance 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 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 material conditions. The mean relatedness between cue and generated error for related materials ($M = .28$, $SD = .09$) was numerically only slightly higher than for unrelated materials ($M = .25$, $SD = .09$), [$t(58) = 1.92$, $SE = .02$, $p = .056$]. To see if participants in the unrelated materials condition altered their guessing strategy as the experiment

progressed, the mean association values for the first 15 items was compared to the last 15 items. There was no difference in the LSA values using this method (M = .24, SD = .09) than the earlier trials (M = .26, SD = .08), [t = 1.09].

	Cue to Error		Target to Error	
Materials	Correct	Incorrect	Correct	Incorrect
Experiment 1a 1b				
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 1c				
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)

Table 1.2
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.

As previously noted, 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. Indeed, as 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 the read long condition was the most beneficial for memory of the target items, and they failed, rather dramatically, to appreciate the benefits of making errors. Furthermore, though performance in each of the three different learning conditions varied greatly between materials, the metacognitive ratings were similar. Comparing materials conditions, it is clear 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 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 1c

The third experiment endeavors to replicate the results of Experiments 1a1b in a within-participants design, in order 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 Experiments 1a1b. In the within-participant 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, then 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 1a1b, 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 native English speaking Columbia students participated for credit. Mean age was 20.7 years (*SD* = 3.3) and 50% of the participants were female.

Design and materials. A 2(materials: related, unrelated) x 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 1a1b were randomly selected for use in the current 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 the previous experiments. 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 current 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-generate 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 Figure 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 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 expected, participants remembered more of the correct targets for the related compared to unrelated materials, [$F(1, 29) = 344.32$, $MSE = .03$, $p < .001$, $\eta_p^2 = .21$]. Though qualified by the interaction, there was a main effect of Learning condition such that error generation lead 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, $SD = 2.84$), [$F(1, 29)=69.18$, $MSE = 14.59$, $p < .001$, $\eta_p^2 = .71$]. 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 slower than in the read short (M = 5.02 s, SD = 2.19) and in the read long conditions (M = 5.08 s, SD = 1.82), [$F(2, 58) = 6.88$, $MSE = 5.31$, $p < .01$, $\eta_p^2 = .19$]. Items in the error generation condition that were answered incorrectly on the final test took longer to produce than items answered correctly, [$F(2, 58) = 5.73$, $MSE = 4.35$, $p < .01$, $\eta_p^2 = .17$]. 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 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 6 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 Figure 3, the error generation condition for related materials was objectively the best condition for retention, (M = 4.38, SD = .87). However, this condition was only given 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 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 ($M_{metacognitive}$ = 2.37 SD = .82, $M_{performance}$= 1.37 SD = .97), [$z = 3.48$, $p < .01$]. No significant differences were found between the

mean metacognitive judgments and performance rankings for the three other cells (related – read-short, unrelated – read short, and unrelated–error generation), [zs < 1].

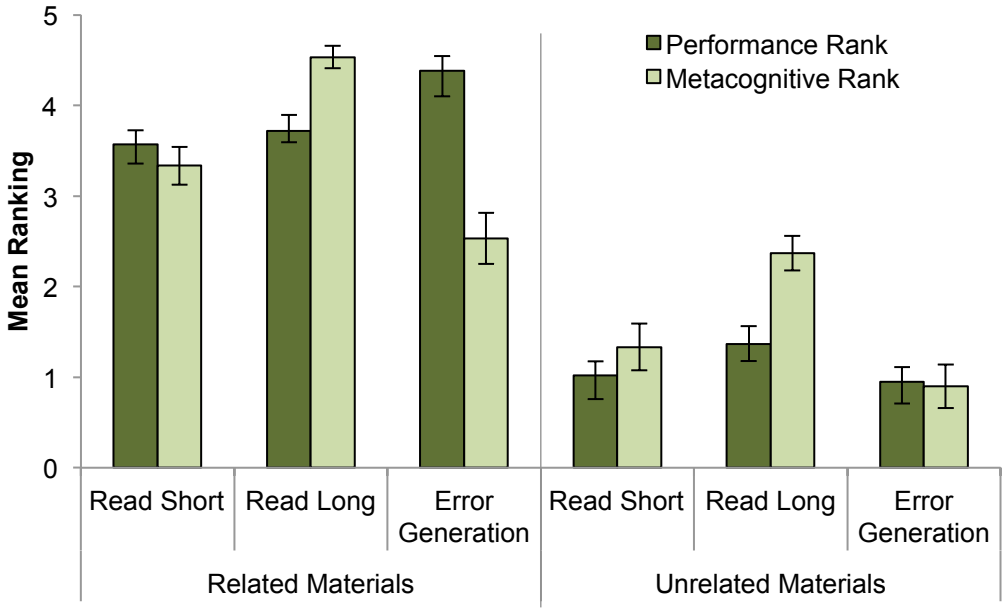


Figure 1.3. **Metacognitive Data for Experiment 1c (within-participants**). Mean ranking of Learning conditions based on correct performance on the final cued recall test and subjective metacognitive judgments. The condition with the highest proportion correct or subjectively rated the best was assigned a score of a 5. Second best was assigned a score of 4 and so forth, while the worst condition was assigned a 0.

Latent semantic analysis. The results from the LSA mirror those of Experiments 1a 1b, 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 = 0.01, p < 001], though 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 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), [$t(29) = 17.51, SE = 0.01, p < .001$].

Discussion

Experiment 1c replicated the findings of Experiments 1a 1b: 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 to 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, then there should have been some benefit from incorrect guessing for the unrelated materials in Experiment 1c. Insofar as items were randomized, 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. This framework would predict that

positive transfer will result if the error and target were 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 (1949) 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. Compared to 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 of the Read long or read short conditions.

Since the time of Osgood, 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 to 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 thereby making the target more meaningful. Though one might engage in elaborative processes for unrelated materials, this elaboration might be in vain. For example, if provided with the word "attack", when one tries to generate a response, one will presumably think about what it means, and erroneously generate "dog".

When the related target, "defend" is displayed, the connection is clear and one can draw a more elaborate and meaningful relationship than simply when one 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 reading as compared to 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. Anderson, 1983). Carpenter (2011) suggests that retrieval helps in activating semantically related information above restudy. In other recent work, Grimadli and Karpicke (2012) found error generation benefits only for semantically related items, a finding consistent with the results presented in the current paper. 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, 1974).

An episodic 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 study found beneficial effects of the episodic link, however, only when it could both be retrieved at time of test, and when it elicited the target item.[3] In the current 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 an episodic 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 an episodic 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. Therefore, it is possible that both of these effects can occur simultaneously.

The RT data are readily interpretable within the *'error as an episodic mediator'* hypothesis. Participants took longer to produce a response on the final test for the error generation condition as compared to the read long and read short conditions. When attempting 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 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 also be assumed to take time. A number of semantic activations models predict longer RTs with a higher number of associated concepts (see ACT-R and Fan-effect: Anderson, 1974; Anderson & Reder, 1999). These models could also predict that participants' RTs would be slower 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." (pg. 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 as global retrospective estimates of performance (GREPs) have been shown in other experimental situations to make use of information acquired during the criterion test to help inform judgments (c.f., Hertzog, Price & Dunlosky, 2008). Retrospective judgments, therefore, have been shown to be more accurate than predictions of performance (see Pieschl, 2009 for a review). For this reason, it is surprising that there seems to be such a large disconnect between subjective performance rankings and actual performance. However, there may be ways to eradicate the metacognitive mismatch if people's attention, at time of test, were more clearly focused on the effect that the various experimental conditions had on memory performance (c.f. Benjamin, 2003; King, 1991; Zimmerman, 2000).

Though currently we cannot make any claims in regards 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 an "ease of processing" heuristic (see Koriat & Ma'yan, 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 & Olsen, in press; Koriat, 1997; Koriat, 2008; Nelson & Dunlosky, 1991; Rawson & Dunlosky, 2002; Rhodes & Castel, 2008). Since error generation might not have seemed as easy as reading the answer (perhaps at both retrieval and at encoding), participants would 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 current 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) generating an error and receiving corrective feedback is better for learning than simply studying. Though 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 one's error and the correct answer.[4]

The second point of interest to educators comes from the metacognitive monitoring results and will be further addressed in Chapter 3. 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 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 the regulation of learning (Finn, 2008; Metcalfe & Finn, 2008; Son, 2004; Son & Kornell, 2008; Son & Metcalfe, 2000; Stone, 2000, also see Metcalfe, 2009). Thus, it seems unlikely that the learner, without further training of his or her metacognition, will implement this highly effective learning strategy.

[4] However, some caution is needed in implementing this recommendation, given that errors may have detrimental effects for memory-impaired individuals, as Clare and Jones (2008) have reviewed. It is not yet known if error generation, in instances where the errors are related to the targets, as studied here, will lead to enhanced or diminished performance for young children or those with learning disabilities.

Chapter 2

Exploring Memorial Mechanisms of the Error Generation Effect

Exploring Memorial Mechanisms of the Error Generation Effect

Mistakes are inevitable. We constantly strive to avoid errors, but are they always harmful for learning? Under certain circumstances, making an error prior to learning the correct answer helps compared to simply studying the correct information (Grimadli & Karpicke, 2012; Hays, Kornell, Bjork; 2012; Huelser & Metcalfe, 2012; Kornell, Hays & Bjork, 2009; Slamecka & Fevreiski, 1983). In Chapter 1, we introduced a standard version of the *error generation paradigm*. In the error generation condition, a cue word is presented on the screen, and one must guess what she thinks the correct answer will be. After producing an incorrect response, the correct target is presented for study. However, in the error-free or "read" condition, both the cue and target are displayed simultaneously, and there is no opportunity to commit an error. On a later memory test, memory for the correct target response is greater after generating errors during learning than reading the word pairs without committing an error (Grimadli & Karpicke, 2012; Hays, Kornell & Bjork, 2012; Huelser & Metcalfe, 2012; Kornell, Hays & Bjork, 2009; Slamecka & Fevreiski, 1983). Additionally, other studies have shown that multiple incorrect retrieval attempts led to better memory for the correct feedback than producing fewer incorrect responses (Arnold & McDermott, 2012; Izawa 1967, 1970).

To date, the learning advantage of error generation seems to be found only when the error is semantically related to the target. In previous research, when the error is unrelated, performance was no better than that in the error free conditions (Grimadli & Karpicke, 2012; Huelser & Metcalfe, 2012, Chapter1). Furthermore, the more semantically similar the error is to the correct answer, the more likely the individual is to give the correct target at final test (Slamecka & Fevreiski, 1983). Though a number of explanations have been put forth as to why generating an error for related materials potentiates learning of the correct response (c.f. Arnold

& McDermott, 2012, Grimadli & Karpicke, 2012; Hays et al., 2012; Huelser & Metcalfe, 2012; Kornell et al., 2009; Slamecka & Fevreiski, 1983), one common thread is the semantic relation or mediation between the error and the correct answer, which we will refer to as the *semantic mediation* hypothesis. The semantic mediation hypothesis concerns the degree to which the error can semantically connect to the correct answer, and this is what determines the effectiveness of the error committed. Based on several semantic network models, the closer two items are in semantic space (i.e. the more related they are), the more activation one item will receive as a result of activation of the other item (Anderson, 1983; Collins & Quillian, 1972; Collins & Loftus, 1974, Neely 1976, Posner & Snyder, 1975). For example, the word "finger" is thought to activate concepts that are related, such as "hand," more than unrelated concepts such as "hillside". This enhanced activation attributable to priming from related information makes an item more likely to be recognized or recalled later due to increased accessibility (Neely, 1976; Posner & Snyder, 1975; also see Higgins, 1996). Therefore, when a person is presented with a prompt such as 'wrist-__?__', producing the error 'finger' should activate concepts related to 'finger' including the concept 'hand.' The activation of the error 'finger' would be expected to make 'hand' more available and more accessible in future memory processing, as compared to the case in which 'finger' was not generated.[5]

Additionally, a semantically related item, which here is the error, can serve as a mediator (prompt, cue, link, or stepping-stone) leading from the cue to the target and enhance memory for the correct target (McKoon & Ratcliff, 1992; and see Carpenter 2011; Pyc & Rawson, 2010).

[5] In addition, by constricting errors by providing the first few letters of the error (e.g. tide – wa____), these related, albeit constrained errors did not lead to better memory above reading (Grimadli & Karpicke, 2012). These results were interpreted to indicate that when generating an error, part of the benefit is derived from "searching" and subsquently activating more related information. Here, by restrictng the search, one does not obtain the benefit of generating an error.

However, an unrelated error should not provide such a stepping-stone. It would either not activate the correct target or it could guide the learner down the wrong semantic path (away from the correct target) and ultimately not benefit recall.

Along similar lines, one can draw a connection to the Gestalt Principle of Good Fit, or *'Prägnanz'* which is based upon regularity, orderliness, uniformity, and degree of coherence (Koffka, 1947; Todrovic, 2008). According to the *associative symmetry hypothesis* of Gestalt Psychology, new items are incorporated into a novel holistic representation (Kahana, Howard & Polyn, 2008), which benefits from a more unified and orderly unit. It could be expected then that unrelated errors should decrease the orderliness of the representation, and hurt memory more than simply studying (See Anderson & Bower, 1973 for more on Gestalt theories of memory). In sum, the data to date are consistent with the Semantic Mediation hypothesis; only semantically related errors, and not unrelated errors, have resulted in enhanced target recall for error generation as compared to error-free study (Grimadli & Karpicke, 2012; Huelser & Metcalfe, 2012).[6]

The Current Paradigm: Polysemous Triplets

In the current chapter, our aim was to explicitly test the semantic mediation hypothesis and investigate the importance of the error-target semantic connection for the error generation effect. To do so, we used materials that led to two different types of errors: semantically related errors, with a direct semantic link between the cue and target, and unrelated errors that led to the

[6] However, in these previous experiments, one important commonality was that the cue and the to-be-remembered target were unrelated, thereby obfuscating the *direct* role of the error-target relationship. In other words, not only was the error unrelated to the target, but the cue was also not semantically linked to the correct target. It could be that generating an error does not provide a memorial benefit if the target is not related to the cue, as a meaningful context might be needed into which to incorporate the target (see Kornell, 2014). Therefore, the resulting design of this experiment begins to examine the role of the error itself.

incorrect semantic domain. If the error generation effect depends upon the error itself being semantically related to the correct answer, we should not see an error generation effect for these incongruent errors. Figure 2.1 provides an illustration of the experimental materials and design.

Participants were instructed to learn word-triplets where the second word of each triplet was polysemous (having more than one meaning, such as PALM). Half of the triplets presented were congruent, so that all three words were in the same semantic space (wrist – PALM – hand). The other half of the triplets was incongruent, so that one meaning was inconsistent with the other two. In other words, the first word of the triplet was of the alternate (non-semantically related) meaning of the polysemous word (*tree* – PALM– hand).[7]

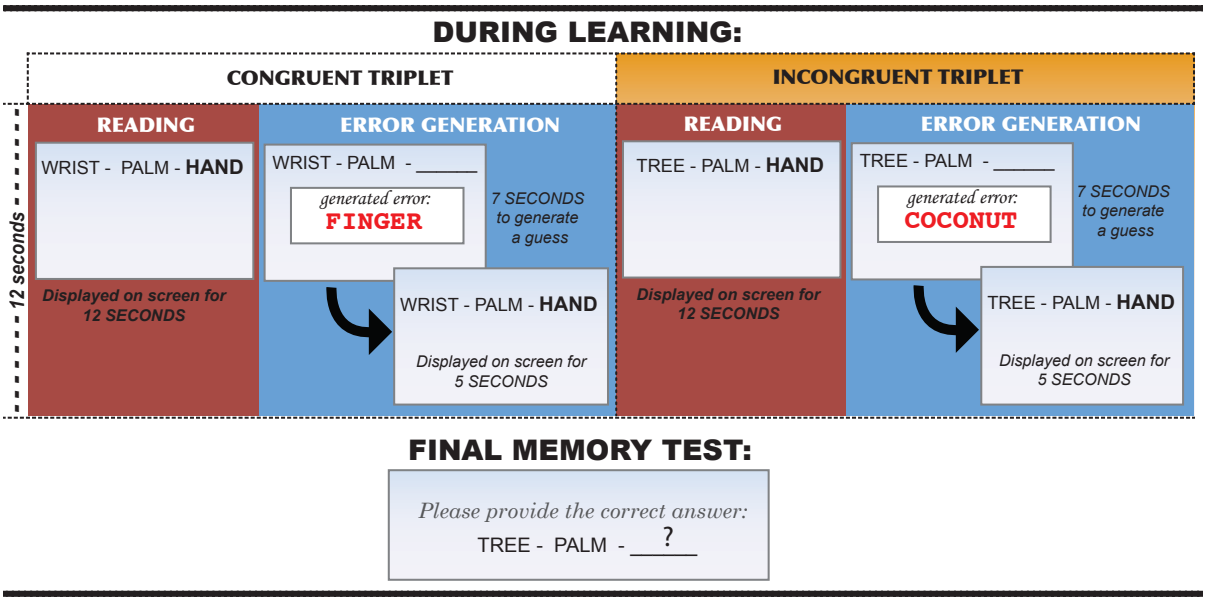


Figure 2.1.

Experimental Materials & Procedural Overview. Example of a Congruent and Incongruent triplet in read and error generation conditions. Both experiments used a 2 (Learning Condition: read, error generation) x 2 (Materials: congruent, incongruent) within-participants design. Each participant studied triplets in each of the four conditions in a random order. Experiment 2b was identical to Experiment 2a, with the exception that participants also provided their original error on the final test.

[7] Participants did not view both the congruent and incongruent versions of each word-pair. The comparison is merely to demonstrate the difference between congruent and incongruent word pairs.

Experiment 2a

Triplets were presented in two different Learning conditions: read or error generate. For read trials, participants viewed the triplet on the screen (e.g. wrist – palm – hand) for 12 seconds. For the error generation condition, participants were given 7 seconds to generate a response to the double-word cue (e.g., wrist – palm-__?__), and then the correct triplet (wrist – palm – hand) was displayed for 5 seconds. If semantic relation is key to the error generation effect, generating an error (such as finger) would be a semantic mediator to the correct target (hand). However, this would not be the case for incongruent triplets (tree – palm-__?__), if the correct answer was again, hand. In this case, however, we expected that people would generate an error consistent with the alternate meaning of palm because it would be that meaning that would be evoked by the concept tree. The incongruent error would lead down the wrong semantic path (such as coconut, sunshine, beach) not to the correct target, 'hand'. If semantic activation or mediation is key, then generating an error in this later case should not enhance memory above reading, and perhaps, lead to decreased learning of the correct response. [8]

Method

Participants. Forty Columbia University undergraduate students (M_{Age} = 22.25, *SD* = 7.3), 62.5% Female, native English speakers) participated for partial fulfillment of course credit.

Materials and Procedure. This was a 2 (materials: congruent, incongruent) x 2 (learning: read, error generation) within-participants design, creating four unique conditions:

[8] This is predicted in part by Marcel (1980)'s priming work which found slower RTs on a lexical deciciosn task for the third word when the triplet was incongruent (e.g. tree-palm-hand) compared to congruent triplets (wrist-palm-hand). This suggests that concepts that are more closely associated with the congruent meaning of the polysemous word receive enhanced activation, and while alternate meanings are suppressed or inhibited (Marcel, 1980).

[congruent: error generation], [congruent: read], [incongruent: error generation], and [incongruent: read].

Polysemous words and related associates were selected from the Nelson, McEvoy and Schreiber's (1998) norms or generated by the author. Triplet creation followed the format used in Marcel (1980), whereby the polysemous word was always the second (middle) word of the triplet, and the target was randomly selected from the associates. Sixty items were randomly selected from the resulting pool of 100 items. Within each set of 60 items, 15 items were then also randomly assigned to sets, which was counterbalanced between learning condition and materials, ensuring items were seen in both congruent and incongruent cases in read and error generation learning conditions between participants. A second random set of 60 items was selected, and also followed the same counterbalancing procedure.[9]

During the learning phase, participants were shown 60 triplets in a random order presented via MediaLab and DirectRT software (Jarvis, 2004), therefore, participants were not aware of what the learning condition (read, error generation) or materials (congruent or incongruent) for each trial. Triplets in the read condition were displayed on the upper left hand side of the screen for 12 s. For error generation, the cue (consisting of the first two words of the triplet) was displayed on the screen for 7 s (WRIST – PALM – ____). During this time, participants typed their prediction for the target (the third word of the triplet). After 7 s, correct feedback (the full triplet, WRIST – PALM – HAND) was displayed for 5s before the next item appeared. Participants read instructions on the screen before beginning this task, and the

[9] We confirmed that the relation between these materials (congruent, incongruent) the polysemous word and congruent or incongruent target semantic relation (using Latent Semantic Analysis, LSA, (Landauer, Foltz, & Laham, 1998) did not differ (*M*s = .23, *SD*s = .02) [$t<1$], and our congruent materials were had a higher degree of relationship between the cue and the target (M =.33, SD =.08) than incongruent items (M = .18, SD = .11) t (59) = 85.56, SE = .02, $p < .001$.

experimenter supervised and explained the procedure during sample trials. Following the learning phase, there was a short (6min) visuospatial filler before the self-paced cued recall test of all 60 triplets presented in a random order.

Results and Discussion

Original errors. Participants produced errors in most of the error generation trials, with slightly more errors for incongruent materials ($M = .95$, $SD = .06$) than congruent ($M = .91$, $SD = .07$), [$t(39) = 2.54$, $SE = .02$, $p < .05$]. Trials in which participants guessed the target correctly were excluded from all further analyses.

We also wanted to ensure that errors on congruent items were related to the target, while for incongruent items, errors would not be semantically related. For example, if the to-be-remembered target was 'HAND', 'finger' might be the error generated for a congruent item, (wrist – palm - __?__). Alternatively, an incongruent triplet (tree–palm- __?__) might lead to the error 'coconut.' Though predicted from priming results using similar materials (Marcel, 1980), using Latent Semantic Analysis (LSA) (Landauer at al. 1998, Landauer & Dumais, 1997) we confirmed that generated errors for the incongruent items ($M = .14$, $SD = .05$) were less related to the target than were the congruent items ($M = .18$, $SD = .05$), [$t(39) = 3.07$, $SE = .01$, $p < .05$]. [10]

Cued recall performance. A 2 (materials: congruent, incongruent) x 2 (learning: reading, error generation) repeated measures ANOVA was computed on the mean correct performance on the final cued recall test. As is shown in Figure 2.2 there was an error generation benefit; generating an error during learning led to improved memory for the target ($M = .65$, $SD = .21$) over simply studying the triplets, ($M = .55$, $SD = .18$), [$F(1, 39) = 16.43$, $MSE = .03$, $p <$

[10] As a manipulation check, we also analyzed the data excluding the trials in which errors were highly related to the target in the incongruent case (one standard deviation above the mean; 11% of error trials). The pattern of results remained the same even with these trials excluded, and therefore, we did not exclude these trials from the current analyses.

.001, $\eta p^2 = .30$]. The benefit of error generation over reading was significant in the congruent condition [$t(39) = 2.64$, $SE = .03$ $p < .02$], as expected from past research, but it was also significantly beneficial in the incongruent condition [$t(39) = 3.97$, $SE = .03$ $p < .001$]. Additionally, overall congruent materials led to higher rates of correct recall on the final test ($M = .64$, $SD = .21$) over incongruent materials ($M = .57$, $SD = .19$), [$F(1, 39) = 12.70$, $MSE = .02$, $p < .01$, $\eta p^2 = .25$].

Original errors were only incorrectly reported as the target response for 9% of the incorrect cued recall responses, and did not differ as a function of materials ($M_{congruent} = .09$, SD=.11; $M_{incongruent} = .09$ $SD = .09$ [$t < 1$],

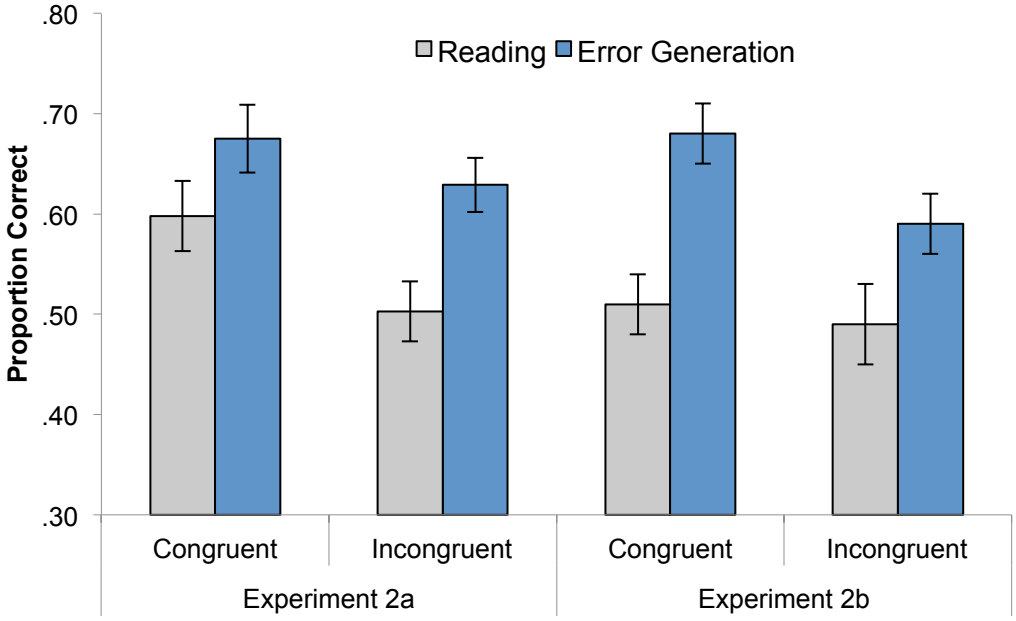


Figure 2.2. **(Experiments 2a and 2b). Cued Recall.** ***Correct*** *performance on final cued recall test as a function of Learning and Materials conditions.*

Error generation during learning led to higher rates of correct target recall for both congruent and incongruent materials than did simply reading the material. *Error bars represent Standard Error of the Mean.*

Discussion

Our results do not wholly support the semantic mediation hypothesis, as we still found an error generation effect for incongruent materials. A close semantic relationship between the error and to-be-remembered information is not the only condition in which generation of errors benefited memory. Therefore, we propose an additional hypothesis to account for these results: Episodic Recollection.

A critical element might be the participants' episodic memory of the original learning event (see Hintzan, 2011; Jacoby & Wahlheim, 2013; Wahlheim & Jacoby, 2013). Previous work has shown that when a meaningful link (either an image, semantic bridge, story, etc.) was generated to connect two unrelated words, the most important predictor for correct recall was the ability to remember one's own generated link (Dunlosky, Hertzog & Powel-Moman, 2005). Perhaps the error is serving as an episodic mediator, linking the cue and the target and bringing the learner back to the original learning episode. For example, when the target is incongruent, a participant might think: "I recollect that my original response to 'tree-palm- __?__' was 'coconut,' but this was incorrect and very different from what the correct answer was supposed to be… I need to go down the other path; PALM can also refer to a part of one's body, so the correct answer is HAND." That said, a similar episodic line-of-thinking might be used for congruent items, though it would result in a straightforward retrieval path. For example, given the cue 'wrist-palm- __?__' the participant might think: "I remember that my original response was 'finger' which had a similar meaning to the correct answer. It was close to correct, but not *quite* the same... The correct answer is HAND." Note that in both the congruent and incongruent instances, getting to the original episode and remembering the error is an important step for producing the correct answer. Furthermore, according to the Episodic Recollection hypothesis, if

there is a breakdown in episodic memory, and one cannot rely upon the original error to guide one to the correct answer, error generation should not be more helpful than errorless-learning.

Experiment 2b

To test the Episodic Recollection hypothesis for the error generation effect, Experiment 2b was identical to Experiment 2a except that in addition to asking participants for the correct answer at final recall, participants were also prompted to provide their original error. If the recollection of the error episode was used in retrieving the correct response regardless of semantic relation to the target, the participant should have also recalled the error itself at the time of target retrieval.

Method

Participants. Forty Columbia University undergraduate students (Mean Age = 19.4 (*SD* = 1.5), 52.5% female, all native English speakers) participated for course credit.

Materials and Procedure. The materials and design were the same as those used in Experiment 2a, though there was a slight procedural difference on the final cued recall test. For each cue, participants provided the correct target *and* their original error if they made one previously. If it had been a reading trial where no original error was made, participants typed "NA" for not applicable. Participants either saw "CORRECT ANSWER?" or "ORIGINAL RESPONSE?" above the cue, depending on the trial type. Order of trial type (being prompted for the correct answer first versus the original response first) was randomized and did not lead to any differences in cued recall performance, [$F<1$].

Results and Discussion

Original errors. Rates of correctly guessing the correct response were similar between materials ($M_{\text{congruent}} = .08, SD = .08$; $M_{\text{incongruent}} = .05, SD = .06$), [$t(39) = 1.55, SE = .02, p = .13$].

These trials are excluded from all other analyses. For the resulting included trials, errors generated during learning are more related to the congruent target (M = .18, SD =.05) than the incongruent target (M = .15, SD =.05), [t (39) = 2.13, SE = .01, p < .05].

Cued recall performance. As in Experiment 2a, we conducted a 2 (materials: congruent, incongruent) x 2 (learning: read, error generation) repeated measures ANOVA on proportion correct on the final cued recall, the means of which are displayed in Figure 2.2. Overall, participants remembered the target more often for congruent items (M = .60, SD = .21) than for incongruent items (M = .53, SD = .22), [F (1, 39) = 10.47, MSE = .01, p = .002, ηp^2 = .21], but importantly, generating errors still resulted in better target recall (M = .64, SD = .21) than simply reading (M = .50, SD = .22), [F (1, 39) = 37.92, MSE = .02, p < .001, ηp^2 = .49]. Furthermore, the benefit for generation of errors over reading was evident for both congruent and incongruent materials [t(39) = 7.12, SE = .02 p < .001, t(39) = 3.36, SE = .02 p < .01], respectively), though slightly larger for the congruent condition, as shown by the significant interaction, [F (1, 39) = 5.24, MSE = .01, p =.028, ηp2 = .12]. As this interaction was not present in Experiment 2a, we compared performance between experiments for each of the four conditions. The only difference trending toward significance between these two experiments is the congruent read condition is slightly higher in Experiment 2a than Experiment 2b [t (78) =1.86, SE = .05, p = .067. All other comparisons are non-significant, [ts <1].

Recall of original errors and conditional analyses. The primary interest was whether or not recollection of errors was crucial for error correction. Overall, participants remembered their original errors on over half of the trials for both congruent (M = .60, SD = .24) and incongruent materials (M = .64, SD = .25), with memory for prior errors being slightly higher in the

incongruent condition, [$t(39) = 2.16, SE = .02, p < 05$]. Overall, 10% of incorrect responses on the final test were original errors ($M_{congruent} = .09, SD = .17, M_{incongruent} = .10, SD = .14$), [$t < 1$].

Still, was target recall higher when the original error was recalled versus not? First we examined proportion correct as a function of if the error was recalled or not for both material types, and these data are presented in Figure 2.3.

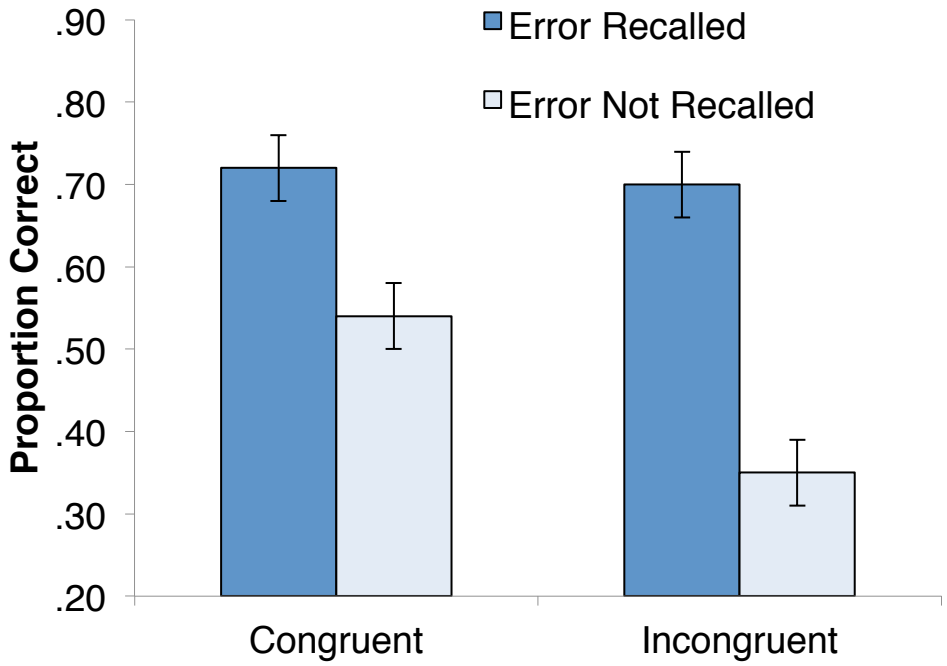


***Figure 2.3*. Conditional Cued Recall**. Cued recall performance conditional upon recall of original error for Congruent and Incongruent items, compared to Read conditions (Experiment 2b).

When the error could be recalled, memory for the correct answer was greater than when the original error was not reproduced on the final test. When the error was not recalled, correct performance was no better than the read condition for congruent items, and even worse than the Reading condition for incongruent items. *Error bars represent Standard Error of the Mean*.

As expected by the episodic recollection hypothesis, when the original error was produced, correct performance on the target recall was 56% greater ($M = .71, SD = .24$) than when the error could not be recalled ($M = .45, SD = .27$) [$F(1, 38) = 33.47, MSE = .01, p =$

.002, ηp^2 = .47]. Though congruent answers were recalled more (M = .63, SD = .27) than the incongruent target responses (M = .53, SD = .25), [F (1, 38) = 10.59, MSE = .04, p < .01, ηp^2 = .22], this effect is qualified by an interaction between Materials and Error recall, [F (1, 38) = 6.09, MSE = .05, p =.018, ηp^2 = .14]. When the error was recalled, correct memory for the target was nearly identical in the congruent (M =.71, SD =.26) and incongruent conditions (M =.70, SD = .23), t < 1). However, when participants could not recall the original error, correct recall of the target was higher for congruent items (M = .54, SD = .28) compared to incongruent items (M = .35, SD = .27), [t(39) = 3.45, SE = .05, p < .01].

Notably, when the errors were not recalled at the final test, there was no benefit from error generation (M = .45, SD = .27) as compared to the read condition (M =.50 SD = .22), F <1. However, not recalling the error in the incongruent condition had more severe consequences than for congruent materials, (significant interaction: [F(1,39) = 7.31, MSE = .05, p = .01, ηp^2 = .16]). When the original error was not recalled for congruent items, performance in the error generation condition (M = .54, SD = .28) was not better than in the read condition (M =.51, SD = .20), [t<1]. However, for incongruent items, error generation (M = .35, SD = .27) was significantly worse than in the read condition (M = .49, SD = .23), [t(39) = 2.74, SE = .05, p < .01].

General Discussion

The findings presented here offer support for the episodic recollection hypothesis: One's error contributes more than simply activating semantically related materials, and the error itself can serve as an episodic memory link. When there is a breakdown in episodic memory and one cannot rely on the original error to guide one to the correct answer, error generation is no more

beneficial than simply studying.[11] These findings are consistent with recent work on *recursive reminding*, as discussed by Hintzman (2004, 2011) and elaborated upon by Jacoby and Wahlheim (2013) (and also Wahlheim & Jacoby, 2013). In Wahlheim and Jacoby (2013) when asked to recall the most recent item of a category, category items that had been previously studied were also reported to come to mind. Furthermore, when the previous item was retrieved along with the more recent item, retrieval was enhanced compared to when no recollection was present. Wahlheim and Jacoby (2013)'s findings support that upon activation of the original episode, other elements associated with this episode are also accessed. If the error itself can be recollected, it should serve as an additional episodic retrieval cue or a reminder of the target. As serial position is also preserved during recursive reminding (e.g., Hintzman, 2011; Jacoby & Wahlheim, 2013) and episodic remembering in general (c.f., Conway, 2009; Howard & Kahana, 2002; Kahana et al. 2008), this episodic memory affords the learner to know which was her error compared to the correct target. In sum, this episodic recollection account of error generation draws upon Tulving's (1983) seminal encoding specificity principle: the context for which a cue was encoded is the optimal retrieval cue. Hence, if the error is aiding one to retrieve the original context or episode, correct memory should be also enhanced by being able to episodically remember your mistake.[12]

---

[11] While we recognize that not recalling an incongruent error leads to worse performance than reading, here, we note that the his difference is perhaps due to a higher probability of correctly guessing the target as compared to guessing the target for (cont) incongruent items as Latent Semantic Analysis predicts. We also recognize an alternate explanation, that even when one cannot recall her original error, the prior activation of related concepts aided in correct recall of congruent targets, yet this activation could not spread to the correct target for incongruent items. Regardless, even if semantic mediation might aid when the error (episodic mediator) is not present, it cannot be the sole mechanism driving the error generation effect.

[12] However, if episodic recollection is key to the error generation benefit, then it might be best to be cautious in unilaterally prescribing error generation as an effective study method. Episodic recollective processes could be problematic for some, in which case, generating errors would not be helpful. (cont.)

In conclusion, this learning benefit of generating errors is not simply a result of how close the semantic connection is between the error and correct answer, and Semantic Mediation cannot be the only explanation for the error generation effect. Support was found for the episodic recollection hypothesis, as recalling the original error did not interfere with the ability to recall the correct response. In fact, error retrieval during test *enhanced* correct recall, even for incongruent materials. This suggests that an error can serve as a mnemonic episodic bridge to the original learning episode. Instead of simply overriding and erasing mistakes, it is important to utilize them as a stepping-stone or an episodic mediator to arrive at the correct answer. To learn from our mistakes, we might need to remember where we went wrong.

In fact, our findings parallel with research on the lack of an error generation effect for those with episodic memory deficits (Baddelely & Wilson, 1993; Clare & Jones, 2008; Hanman & Squire, 1995; Komatsu et al., 2000; Kalla, Downes, & Van den Brock, 2001; Tailby & Halsam, 2005). Such individual differences based upon the episodic recollection hypotheses of error generation are of great importance.

Chapter 3

Metacognition of Error Generation: Stable or malleable bias?

Metacognition of Error Generation: Stable or malleable bias?

What strategies are best for learning? Although it has been shown that knowledge about strategies for learning increases from childhood to adulthood, it is clear that adults are not always aware of which strategies lead to the relative best performance (see Bjork, Dunlosky & Kornell, 2013 and Veenman, 2010 for reviews). This understanding or knowledge of a study strategy's effectiveness is a critical aspect of the broad-reaching construct of *metacognition*, which can be defined as thinking or reflecting upon one's own thoughts or abilities (Flavel, 1979; Nelson & Narens, 1990; Metcalfe & Dunlosky, 2009).

The current question of interest regarding the error generation effect is a metacognitive one: *Are learners able to accurately metacognitively monitor (detect/assess) that making an error or guessing incorrectly can be an extremely effective study tool?* In Chapter 1 (Huelser & Metcalfe, 2012), we demonstrated a metacognitive illusion, as moments after completing the criterion test, participants failed to rate error generation as best for performance. These findings were surprising, as postdictions (retrospective) have been shown to be less difficult than predictions (prospective) (c.f., Hacker, Bol, Horgan, & Rakow, 2000; also see Pieschel, 2009 for a review). Therefore, the goals of the current chapter are to begin to exclude some possible explanations of the bias against error generation as an effective learning strategy, in addition to investigating if this bias is stable or if it can be correctly updated.

Underconfidence in Error Generation as an Effective Learning Strategy

In Huelser and Metcalfe (2012), learners failed to accurately assess strategy effectiveness on an aggregate, or global level, which is one key component of many metacognitive models (Nelson & Narens, 1990, Hertzog, Price & Dunlosky, 2008; Schraw, 1994; Schraw & Moshman,

1995; Winne, 1996; Winne & Hadwin, 1998; Zimmerman, 2000).[13] Yet, how does one monitor performance, especially in aggregate? It has been suggested that learners do not only use task specific cues when making metacognitive judgments, but also rely on more stable metacognitive knowledge (Flavel, 1979; Winne, 1996), particularly at the global level (c.f. Hertzog, Price & Dunlosky, 2008). As metacognitive knowledge can be influenced by motivation, task-difficulty, and self-knowledge (Hertzog et al. 2008; Flavel, 1979), one potential source of bias may be based upon negative apriori (pre-experimental) beliefs about errors. Despite doing well on the final test, learners might be biased that making errors is negative, and therefore, overrule or fail to update their metacognitive knowledge about errors and learning. Studies have shown aversive effect associated with making an error (e.g. Hajcak & Foti, 2008), and "[t]he experience of failure is associated with negative emotion, lowered self-esteem, reduced intrinsic motivation, and lower expectancies of future success, particularly when the failure is attributed to internal causes." Chase, 2012, pg. 2). Furthermore, on a measure of error attitudes (Error Orientation Questionnaire), errors were associated with guilt and fear (Rybowiak, Garst, Frese, & Batinic, 1999). Therefore, it is not surprising one might believe that errors are harmful for learning. If one experiences the errors as failures and is unable to attend to the corrective feedback as a result of distress or nervousness due to making the errors, then error generation might not be beneficial for learning (Zhao, 2011). Even within the classroom, Tulis (2013) recently showed that teachers

[13] There are different methods to measure global monitoring in Huelser and Metcalfe (2012) participants simply ranked learning conditions. However, subjective strategy preference can also be assessed by estimates of how many items (or a proportion of items) are answered correctly. In the current set of experiments, we refer to these global performance assessments as Global Retrospective Estimates of Performance (GREPs). The terminology in the literature is mixed; several have used,"Global Differentiated Postdiction" (c.f. Dunklosy & Hertzog, 2000, Hertzog et al. 2008; Schraw, 1994), while others in the Self-Regulated Learning Literature also refer to item-monitoring as cognitive evaluations and global as metacognitive evaluations (Winne & Hadwin, 1998; Greene & Azevedo, 2007) Others have used the terms "on-line" versus "off-line" respectively (Van Hout-Wolters, 2000; Veenman, 2005).

rarely encouraged error risk taking and were not likely to suggest mistakes as opportunities for learning.

Despite the potential pre-experimental incorrect metacognitive knowledge that might contribute to the underconfidence effect, we still are not sure if the inaccuracy is only at the global level (e.g. How did I perform on this test, overall?), or also present during item-by-item monitoring (e.g. Did I answer this question correctly?). These item-specific retrospective confidence judgments are also a critical aspect of metamemory models as a form of metacognitive monitoring at retrieval (Nelson & Narens, 1990; Dunlosky, Serra & Baker, 2007; Hertzog et al., 2008). There are several experiments, spanning various materials, which suggest learners often excel at making item-level confidence judgments with considerable accuracy (Butterfield & Metcalfe, 2001; Dunlosky & Hertzog, 2000; Hertzog et al, 2008; Higham, 2002; Matvey, Dunlosky, Shaw, Parks, & Hertzog, 2002; Schraw, 1994; Veenman, 2010). However, if metacognitive monitoring is inaccurate on an item-level, then it is likely that global metacognitive monitoring would also be inaccurate, especially if item-by-item judgments are used to inform global judgments (as seen in Dunlosky & Hertzog 2000; Hertzog et al., 2008). Put differently, if the learner (falsely) believes she is answering several of the items from the error generation condition incorrectly, we would expected her to be underconfident when assessing the effectiveness of error generation on later memory performance. A critical first step in understanding why learners are not aware of the error generation benefit is to measure confidence of performance accuracy on an item-by-item level; if the underconfidence lies here, it is likely also to lie at the global level.

Beyond simply understanding on which level—item-by-item and/or global—underconfidence occurs, it is of great research interest to explore possible mechanisms by which

underconfidence can be reduced or eliminated. In other words, can one discover that generating errors can be beneficial for memory *without* explicitly being told that her performance was best in the error generation condition? This is an important question, as Veenman (2010) reiterates from a historic overview that 96% of metacognitive instructions in the classroom are implicit. Teachers did not explicitly describe the metacognitive strategies used in the classroom, nor why they were beneficial for learning (Veenman, deHaan & Dignath, 2009).

There is evidence both to support and contradict that implicit updating of metacognitive knowledge is possible. Kornell and Bjork (2009) have shown a robust metacognitive "stability bias", such that learners often fail to adjust their metacognitive knowledge to correctly predict future performance. This view is also supported by Winne and Hadwin's (1998) model of self-regulated learning, in which stable metacognitive beliefs are resistant to updating based on performance feedback (see also Greene & Azevedo, 2007). Additionally, it has been suggested that due to high cognitive load demands, such complex content knowledge is not possible to update without considerable guidance (Sweller, Ayres, & Kalyuga, 2011), or unless learners are explicitly informed of which strategy was best for learning (Tullis, Finely, & Benahamin, 2013). Overall, even if learners are accurate on an item-level, on a global level a learner might not be able to retrospectively aggregate her performance on various items while also remembering the learning condition in which each of the items originally appeared (Gigerenzer, Hoffrage & Kleinbolting,1991; Schraw & Niefeld, 1998; Winne, 1996). Faced with such a demanding task, learners might be prone to various sources of bias in making global judgments (e.g. Kahneman, 2003; Manis, Shedler, Jonides, & Nelson, 1993).

However, other work has shown that metacognitive calibration between performance and monitoring can be improved by having learners provide arguments against their answers (Koriat,

Lichtenstien and Fischhoff, 1980). For example, by providing grade-point incentives (Schraw, Potenza & Nebelsick-Gullett, 1993), offering financial incentives (Epley & Gilovich, 2005) and even by guided long-term training over the course of the semester (Nietfiled, Cao & Osborne, 2006) studies have shown that metacognitive knowledge can be updated. Thiede and Anderson (2003) also showed that summarizing after a delay enhanced metacomprehension accuracy. In addition, Dunlosky and Hertzog (2000) and Benjamin (2003) both demonstrated that learners can become more accurate between performance and future learning predictions at the global level without explicit guidance (also see Hertzog et al., 2008). After taking a test, participants made more accurate global predictions about which group of materials (e.g. high frequency vs. low frequency words, Benjamin, 2003) or which study strategies (e.g. Imagery vs. Rote Rehearsal, Dunlosky & Hertzog, 2000) would lead to the best performance. Critically, in Dunlosky and Hertzog (2000) the test itself might not have led to these updated predictions; participants *also* monitored performance on the criterion test by making item-by-item confidence judgments. The subsequent global predictions might not have been as accurate without item-level monitoring. Benjamin (2003) further provides evidence of the importance of item-by-item confidence judgments for updating metacognitive knowledge by manipulating overt item-level performance monitoring, which led to more accurate recognition predictions for a *second* set of words. When participants did not make the item-level judgments during the first set of words, their predictions for the second set remained incorrect. A critical point of updating metacognitive knowledge seems to be on an item-level during the criterion test. As these item-by-item judgments have aided in calibration for other materials and study strategies, perhaps this fine-grained monitoring will aid in updating metacognitive knowledge regarding the effectiveness of error generation.

Experiment 3a

In the present experiment, we address the following two questions: 1) Are learners accurate in their item-level monitoring of performance? 2) Does this item-level monitoring have consequences for strategy knowledge updating? The procedure was nearly identical to Huelser and Metcalfe (2012), except that in addition to attempting to provide target answers during the final cued recall test, half of the participants assessed accuracy for each item on a 0-100 confidence judgment scale (confidence monitoring condition). Those in the control group (no-monitoring condition), provided the correct answer, without overtly making item-level confidence judgments. After the cued recall test, participants gave a global retrospective estimate of performance (GREP), for which they estimated the proportion of items they answered correctly for each of the three learning conditions (read short, read long, and error generation).

By having half of the participants make item-by-item confidence judgments for each item on the final test, we can determine if participants' underconfidence of the memorial benefit of making errors is driven by an inability to know *on an item-level* which questions are being answered correctly (or incorrectly). If participants are able to make accurate item-by-item judgments, but are inaccurate on a global level, we can infer the metacognitive breakdown is not at the item-level monitoring for performance, but instead occurs sometime during the transition from item-level to global. In addition, monitoring might have implications for one's overall assessment of error generation as a learning strategy. If participants are aware of accuracy on an item level, our manipulation of overt-performance monitoring could lead to enhanced global calibration. If so, this would indicate that simply assessing performance for each item aids in updating metacognitive knowledge of error generation as an effective learning strategy.

Method

Participants. Thirty-six Columbia University students participated to partially fulfill a class requirement (77% female, *M* age = 21.14, *SD* = 6.90). Half of the participants were randomly assigned to the control group (no item-level monitoring at test) and the remaining half to the confidence monitoring group (those who made item-by-item confidence assessments).

Materials and design. Sixty weakly associated word-pairs (.05-.054 forward associative strength) from Nelson, McEvoy and Schreiber's (1998) were randomly selected from Huelser & Metcalfe (2012). Monitoring on the final cued recall was manipulated between participants, while learning condition was manipulated within participants, resulting in a 2 (monitoring group: no monitoring (control), confidence monitoring) x 3 (learning condition: read short, read long, error generation) mixed design. Materials were randomly assigned to sets and counterbalanced within learning condition, ensuring items were distributed among the three within-participant learning conditions equally.

Procedure. Participants began the experiment by completing sample trials of each of the three learning conditions. For read short, the intact word pair (cue-target) was displayed on the screen for 5 s, while for read long, the intact word pair was displayed for 10 s. For error generation, the first word of the word pair (cue) was displayed for 5 s with a textbox below, into which participants typed a guess of what they thought the correct second word (target) would be. Following 5 s, the correct word pairing (cue-target) was displayed for 5 s. Participants were instructed to remember the correct target ("what the computer told them") for the later test, not their response. In addition to all instructions on the computer screen, the experimenter verbally confirmed the instructions with the participant, and added, "The task is very difficult. Regardless of what you type, please really try your best to remember the correct second word for later." [14]

[14] This is briefly discussed in Chapter 4.

Learning condition items of read short, read long and error generation were randomized (not blocked). Following the Learning Phase, there was a 6 minute set of visuospatial tasks consisting of assorted puzzles.

The Cued Recall Phase differed slightly depending on the between participants condition assignment. In the no monitoring (control) condition, participants were instructed to type the correct target in response to the displayed cue on the screen (self-paced, no time restriction). To move to the next trial, participants hit the "enter" key, and there was an interstimulus interval of 750 ms before the next cue was displayed. Cues for all 60 items were displayed in a random order. The procedure was nearly identical for the confidence monitoring condition, though in addition to providing the target, participants also made a confidence judgment of the accuracy of their response using a 0-100 Scale [0 = Sure Incorrect, 100 = Sure Correct]. Participants read instructions explaining the additional task prior to the cued recall test. After attempting to provide the correct target (self-paced), the cue remained on the screen with the additional prompt "Confidence? 0-100" displayed above. Participants typed their corresponding confidence judgment (self-paced) before the following cue was displayed. No performance feedback was given.

Following the Cued Recall Test, all participants made a Global Retrospective Estimate of Performance (GREP) on the final cued recall test for each of the three learning conditions. Participants were asked to indicate what percentage (0-100%) of items they answered correctly for items previously displayed in each of the three learning conditions: *"In the first part of the experiment, you studied word pairs in three different ways. Sometimes both words were displayed on the screen together for 5s (Together Short), other times both words were displayed together for 10 seconds (Together Long). In other instances, the first word was displayed*

separately from the second word (and you typed a response) before the correct pairing was displayed (Separate). Think about how well you remembered the correct second word for each of these three learning conditions. Please enter the percentage you think you answered correctly within each of these three different learning conditions." The conditions were renamed "Together Short, Together Long, and Separate" to avoid usage of "Error Generation" due to concerns that the word "error" might bias participants' responses. Following estimates of performance, participants were also asked to briefly explain why they believe they did the best in the condition they specified.

Results

Initial test correct guessing. Participants correctly guessed the target on 3% of the error generation trials (*SD* = .04), and there was no difference between participant groups [t (34) = 1.34, *SE* = .01, p = .19]. These correct trials were excluded from further analyses.

Cued recall performance. As is shown in Figure 3.1, there was a main effect of learning condition, [F (2,68) = 20.61, M*SE* = .01, $p < .001$, $\eta_p^2 = .38$] such that participants remembered

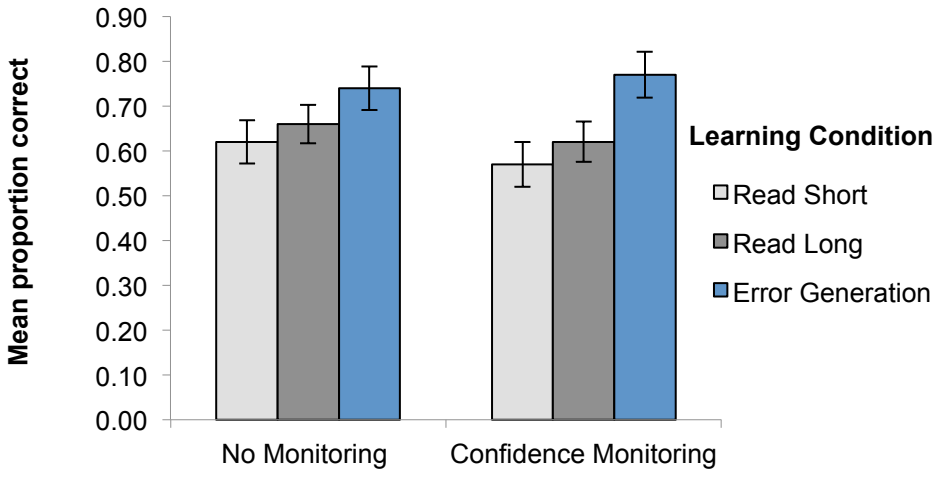


Figure 3.1. (Experiment 3a) **Cued Recall.**
Mean proportion correct on cued recall as a function of item monitoring condition during cued recall (between participants: no monitoring, confidence monitoring) and learning condition (within participants). There was a significant error generation effect for both groups, and this did not interact with monitoring group.

the correct target more often in the error generation condition (M = .78, SD = .19) than the two read conditions ($M_{ReadLong}$ = .65, SD = .18, $M_{ReadShort}$ =.60, SD = .19). Post-hoc tests showed that error generation resulted in higher performance than read long [$t(35) = 5.86$, SEM = .03, $p <$ .001] and read short [$t(35) = 4.03$, SEM = .03, $p < .001$] and that read long resulted in slightly better performance than read short recall [$t(35) = 2.07$, SE = .02, p = .046]. There was no effect of, or interaction with monitoring condition [$F < 1$, F (2, 68) = 1.54, MSE = .01, p = .23, η_p^2 = .04].

Item-by-item confidence judgments. The item-by-item confidence ratings provide insight to whether participants are aware of their performance on an item-level. We only present item confidence ratings for the confidence monitoring condition, as the no monitoring condition (control) did not make these ratings. These means are displayed in Figure 3.2. First we will assess calibration, or absolute accuracy between performance and item-confidence judgments. For each participant, the mean confidence ratings for all trials within subject learning condition

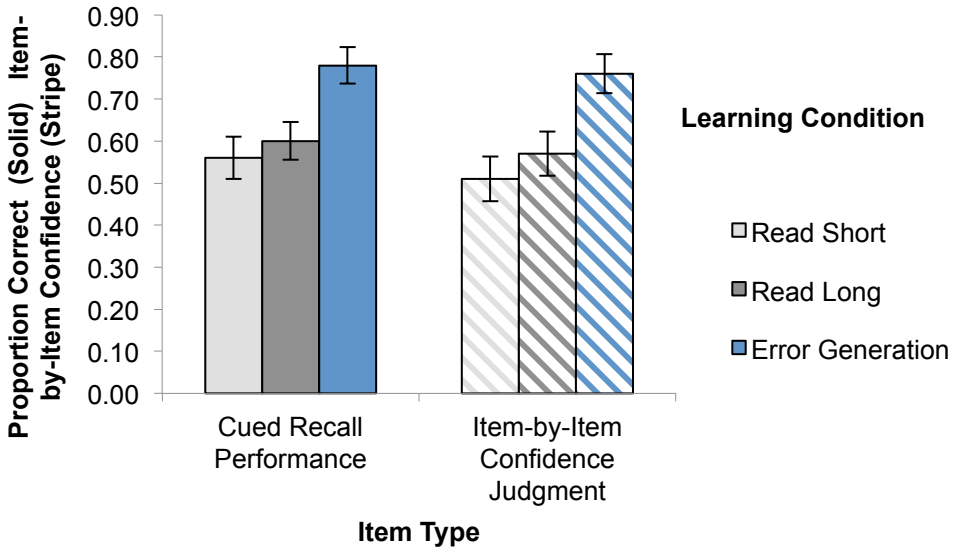


***Figure 3.2. (Experiment 3a: Confidence Monitoring Group*). Cued Recall and item-by-item scores.** Mean proportion correct on cued recall (left panel, solid bars) compared to mean item-by-item confidence judgments (right panel, striped bars). Item-by-item judgments were similar to cued recall performance within each of the three learning conditions.

was computed, and subsequently compared to the mean accuracy of target recall for each of the three conditions. The pattern of mean item-by-item confidence ratings is very similar to that of actual correct performance. There was a main effect of learning condition on item-by-item confidence [F (2, 34)= 24.31, MSE = .13, p < .001, ηp^2= .59], such that participants had the highest mean confidence in the error generation condition (M = .76, SD = .18), followed by read long (M = .57, SD = .21) and read short (M = .51, SD = .22), mirroring the main effect of learning condition from cued recall performance patterns.

Correlations between performance and item-by-item confidence. We also conducted Pearson correlations to see how participants' cued recall score was related to mean confidence across subjects for each of the three learning conditions. These correlations confirmed that higher mean confidence ratings for each learning condition were positively correlated with cued recall performance [$r_{\text{read short}}$ (18) = .90; $r_{\text{read long}}$ (18) = .92; $r_{\text{error generation}}$ (18) = .91, ps < .001].

Learners also knew specifically which items they answered correctly or incorrectly. Relative accuracy was assessed by computing a gamma correlation for each participant. A perfect gamma correlation score of +1.0 corresponds to a perfect positive relationship where all high confidence items are answered correctly, and all low confidence items are answered incorrectly (see Nelson (1984) for further discussion of gamma analysis). The overall mean gamma was .88 (SD = .14). This correlation did not differ as a function of learning condition, (M gamma $_{\text{read short}}$ = .87, SD = .22, M gamma $_{\text{read long}}$ = .83 SD = .17, M gamma $_{\text{error generation}}$ = .90, SD = .11), [F < 1].

Global Retrospective Estimates of Performance (GREPs). GREPs, the mean ratings people gave after completing the task of how well they had learned in each condition, are presented in Figure 3.3. A 2 (monitoring group) x 3 (learning condition) mixed ANOVA

revealed an interaction between learning condition and monitoring group [$F(2,68) = 5.99$, MSE = 03, $p < .01$, $\eta p^2 = .15$]. As was found in our previous experiments (see Chapter 1), when there was no monitoring, read long was judged to lead to the best cued recall learning ($M = .61$, $SD = .20$) over read short ($M = .47$, $SD = .22$) and error generation ($M = .42$, $SD = .23$) [$t\,(17) = 3.25$, $SE = .04$, $p = .005$, $t\,(17) = 2.83$, $SE = .07$, $p = .011$, respectively]. In contrast, in the crucial condition in which participants made item-by-item confidence judgments, the pattern of GREPs was different. In this case, they believed that error generation enhanced cued recall performance. Their GREPs were highest in this condition, and were significantly greater than in the read short condition ($M = .63$, $SD = .26$) [$t\,(17) = 2.51$, $SE = .07$, $p = .023$] though the difference between the read long and error generation GREPs was not significant [$t = 1$].

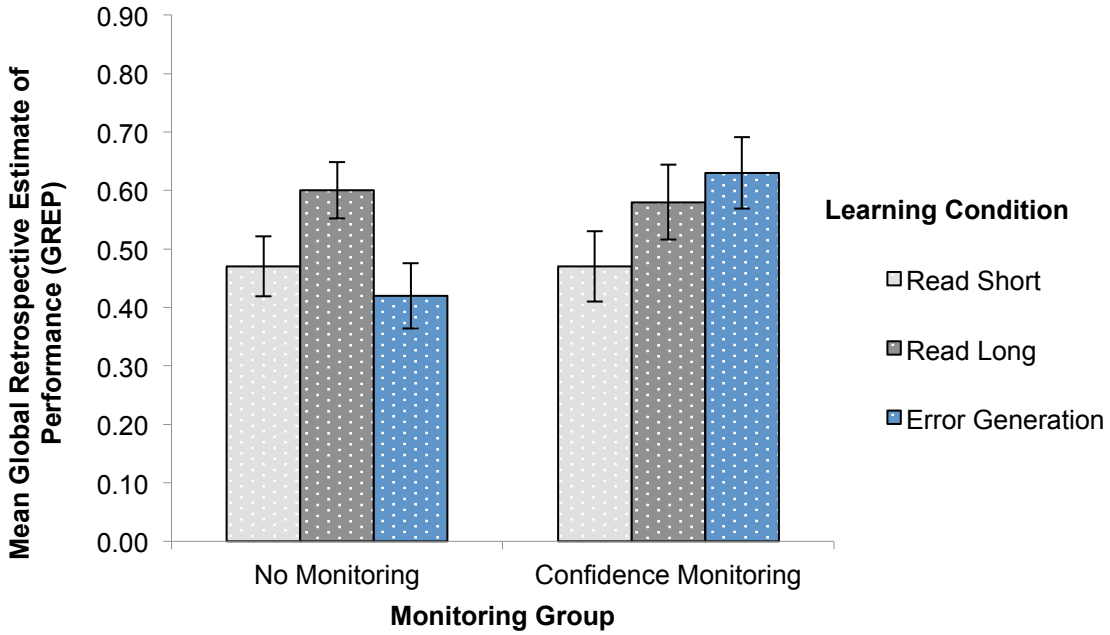


Figure 3.3 (Experiment 3a). **Global Retrospective Estimates of Performance.** Mean Global Retrospective Estimate of Performance (GREPs) as a function of monitoring group (between participants) and learning condition (within participants). Participants in the confidence monitoring group gave higher mean GREPs for the error generation condition than

While GREPs were nearly identical between groups for read short ($M_{\text{no monitoring}}$ = .47, *SD* = .22; $M_{\text{confidence}}$ = .47, *SD* = .25) and read long ($M_{\text{no monitoring}}$ = .61, *SD* = .20; $M_{\text{confidence}}$ = .58, *SD* = .25) [*ts* < 1], in the error generate condition, participants in the control condition estimated performance to be .21 worse than those who made item-by-item confidence judgments ($M_{\text{no monitoring}}$ = .42, SD = .23; $M_{\text{confidence}}$ = .63, *SD* = .26) [*t* (34) = 2.27, *SE* = .08, *p* = .013, *d* = .86]. [15]

Discussion

Those in the confidence monitoring group estimated performance to be better in the error generation condition than did those in the no monitoring control group, who did not make any item-level confidence ratings during the final test. Also important to note is that the item-by-item confidence judgments themselves were *very* well calibrated with actual correct performance, demonstrating that participants knew on an item-by-item basis whether an item was correct or incorrect.

Therefore, even without providing explicit performance feedback, item-by-item confidence judgments reduced the .35 difference between performance and global estimates found in the control group, to only underestimating the effectiveness of the error generation condition by .15 in the confidence group. It seems surprising that typing a number from "0-100" would have such a substantial effect, as one might assume that a test taker would implicitly perform such an assessment even when not explicitly asked to do so. Additionally, previous research has shown that the act of taking a test itself has the potential to update metacognitive

[15] Overall, participants retrospectively rated their performance as best in the read long condition (*M* = .59, SD = .23), followed by error generation (*M* = .52, *SD* = .26) and then read short (*M* =.49 SD = .22) [main effect of learning condition: *F* (2, 68) = 5.13, M*SE* = .03, *p* < .01 ηp^2= .13]. Between confidence monitoring and no monitoring groups, mean overall estimates were similar, signifying that monitoring did not lead to differences in overall global estimates of performance [*F*<1].

knowledge (c.f., Koriat & Bjork, 2006), but in this experiment (and in Huelser and Metcalfe, 2012), we found that the cued recall test alone was not sufficient. While Experiment 3a demonstrated that a reduction in underconfidence is possible, the bias was not eliminated. In Experiment 3b, we sought to further reduce the error generation underconfidence effect.

Experiment 3b

One possible explanation for why underconfidence was not eliminated in Experiment 3a is that participants might not remember *how* each item had been learned. If participants incorrectly attribute most of the incorrectly answered items to the error generation condition, it should not be a surprise that *globally* this learning strategy is perceived as worse for learning.

Therefore in Experiment 3b, the aim was to assure that participants could correctly identify the original learning strategy of an item during the criterion test. Half of the participants indicated in which learning condition (read short, read long, or error generation) an item had been studied during the learning phase. Remembering the original learning condition is a type of source monitoring, as well as an important component of metacognitive monitoring (Dunlosky & Metcalfe, 2009; Nelson & Narens, 1990). Given that we found reduced underconfidence on global ratings due to item-level confidence monitoring, we predicted a similar effect given source monitoring (c.f., McCloskey and Zaragoza, 1985). If monitoring accuracy on an item-level updates metacognitive knowledge, heightening awareness of which items are from which conditions should also aid in reducing the metacognitive bias. Furthermore, what if one reflects upon both how one is performing and how an item was learned? A greater reduction of the bias might be possible in this case, while being aware of only source or only accuracy might only partially update metacognitive knowledge.

However, even if participants can identify the correct source at the item-level, this additional source monitoring still might not enhance global calibration above item-level confidence monitoring. Again, retrospective global measures have not been as predictive as those that occur during the task (Veenman, 2010). This could be in part due to the greater complexity of global judgments (Hacker, Bol & Keener, 2000). Given that our experiment consisted of 60 tested items, with the order of learning conditions randomized during learning and again at test, mental summation is not likely (Winne, 1996). Tullis, Finley and Benjamin's (2013) findings also suggest that further reducing the error generation underconfidence bias is not likely, even when prompted to monitor source. They recently demonstrated an inability to update metacognitive knowledge about the effectiveness of the learning conditions without directed guidance. Predictions for performance on a subsequent list were most accurate when participants received *explicit* feedback on their global performance for each of the learning conditions (testing versus study) from the previous list (Experiment 4). Specifically, participants were told how many items (out of 16) they had answered correctly for each of the two learning conditions. When participants were only informed of accuracy and learning condition source on an item level (Experiments 2-3), participants did not make more accurate predictions for future performance for tested items. Therefore, even if one knows on an item-level both performance accuracy and the type of learning condition, this might not be sufficient to completely override the error generation underconfidence effect.

Metacognitive control: implications of monitoring on future strategy selection. Of additional concern are the implications of inaccurate global assessments of performance. Does incorrect global monitoring have consequences for later study strategy selection? This question is of great importance, as monitoring has consequences for metacognitive control or regulation

(Metcalfe, 2002, 2009; Metcalfe & Kornell, 2003, 2005; Metcalfe & Finn, 2008; Son & Kornell 2008; Son & Metcalfe, 2000; Son, 2004; Son & Sethi, 2006; Stone, 2000). For example, items judged to be either too easy or too difficult to learn are not as likely to be selected for restudy, as easy items are already learned (hence, no need to re-study) and items that are too difficult might simply not be learned. (See Metcalfe, 2009 for a brief review research on metacognitive monitoring and control). In a clever set of experiments, Metcalfe and Finn (2008) were able to isolate the effect of monitoring on control by demonstrating that monitoring had consequences for metacognitive control simply by manipulating *perceived* learning of word pairs, without altering *actual* performance. More items were chosen for restudy when items were judged as less well learned, even though performance did not differ. Given the potential for monitoring to lead to enhanced metacognitive control, we wanted to explore if monitoring on an item level had consequences for strategy selection. Would learners be more likely to select error generation as a study strategy in the future?

Though one might assume enhanced monitoring should automatically lead to better strategy selection, there are reasons to predict a disconnect between these two elements of metacognition. Strategy selection might be influenced by other factors, not simply related to performance monitoring or judgments of performance: "According to metacognitive theories, beliefs about the task, oneself, and the repertoire of strategies one has available all can influence initial strategy selection, and whether an individual continues a strategic approach or alters it in the face of performance-goal discrepancies." (Hertzog et al., 2008, pg. 430) Furthermore, perceived difficulty of the strategy, not just its effectiveness, could also influence whether a particular strategy is chosen for later use (Rabinowitz, Feeman & Cohen, 1992). In addition, Kornell and Son (2009) noted a disconnect between monitoring and control where participants

reported taking a test to lead to worse performance than repeated studying, yet they still selected to be tested instead of study for a subsequent to-be-learned list. Therefore, there are varying predictions of whether item-level source monitoring will lead to enhanced GREPs, and consequently a preference in choosing error generation as a study strategy for a future test. Even if participants give higher GREPs for error generation post-item-level monitoring (and realize it is helpful for learning), they still might not chose this strategy in the future, especially if one might perceive making errors in a negative light. Therefore, in the following experiment, we sought to address this question of metacognitive control by asking participants to select how many read short, read long, or error generation items they would like to study for a future test.

To summarize the research questions of Experiment 3b we investigated the following: 1) Was underconfidence in error generation as a learning strategy simply due to an inability to correctly monitor which items were learned in the error generation condition? 2) With an additional source monitoring task, could we enhance calibration and possibly eliminate the error generation underconfidence effect? 3) Did estimates of performance in the three conditions translate into corresponding study choices? 4) What performance consequences would ensue if people actually used their stated condition preferences as their study strategy?

Method

Participants. Sixty Columbia University students participated for partial fulfillment of a class requirement (M_{Age} = 21.06 years (*SD* = 5.67), 60% Female). Participants were randomly assigned to the four between subject monitoring conditions.

Design and Materials. We used a mixed factorial design: 2 (source monitoring: no, yes) x 2 (confidence monitoring: no, yes) between participants factors, and 3 (learning condition: read short, read long, error generate) within participants factors. Each participant answered questions

in each of the three learning conditions. However, by crossing confidence and source monitoring between participants, we had four unique between participant conditions: 1) no monitoring (control, no source + no confidence), 2) confidence monitoring only, 3) source monitoring only, 4) confidence + source monitoring.

Conditions of source monitoring only and confidence + source monitoring are novel to the current experiment, while no monitoring and confidence monitoring only correspond to those from Experiment 3a. The same materials from Experiment 3a were used for the current design.

Procedure. Procedurally, the learning phase and visuospatial filler were identical to those used in Experiment 3a. For the two new between-participants conditions involving source judgments, participants were given instructions prior to the start of the final test. On source judgments trials (conditions 3 and 4), participants indicated in which learning condition the cue had originally appeared "in the first part of the experiment." Only the CUE word (*not the target)* was displayed and "Original Presentation?" was above the cue. Directly above the text box, the following instructions were provided: "Please label in which condition this word was presented. If from the Separate condition, please write your previous answer." The following prompts were provided below the text box on each trial as reminders of the three learning condition options: TS = Together Short, TL = Together Long, Original Response = Separate. [Footnote: We assumed that if one could provide their original response, this would be an appropriate analog for knowing the source was the error generation condition. On some trials, if participants could not remember their original error, but knew it was an error generation trial, they wrote "S" or "Separate" indicating the correct source. These were coded as correct.] The source judgment trial for each cue occurred after the cued recall trial for that item. For the source + confidence group, the order

of trials for each item was the following: 1) cued recall, 2) item-by-item confidence rating, 3) item-by-item source judgment. All trials were self-paced.[16]

The Global Estimates of Performance (GREPs) were identical in each of the four between final test ratings conditions. Participants made GREPs by providing estimates of their overall average performance on the final test for each of the three learning conditions. After giving ratings of performance, participants briefly explained by typing into a text box why they estimated that condition to be the best for learning. This was free response and self-paced.

Following GREP assessments and explanations, participants indicated how many items they would like to study in each of the three conditions by making a Study Strategy Choice. Participants were told that there were an additional 20 cue-target pairs to learn, and that they would be tested on these items (though this did not occur). They specified how many items (out of the 20) they wanted to allocate between the three different learning conditions in order to maximize performance on a later test (again, which never occurred). Subsequently, they explained their reasoning for why they selected the most items in their preferred condition. Participants were fully debriefed and thanked for their participation.

Results

Initial test correct guessing. Participants correctly guessed the target for 3% of the learning phase trials (SD = .04), and there was no difference between participant groups [$Fs < 1$]. These correct trials were excluded from further analyses.

[16] There was a slight difference in ISI intervals between experiments. The control condition here was 1000 ms, and the two single monitoring conditions (source only, confidence only) had ISIs of 500 ms. Total Trials Times are reported in Appendix A for completeness, though not further addressed here. Critically, there were no differences in RTs to produce the target across the between subjects monitoring conditions for both Experiments [$Fs<1$]. However, on total trial time (even when correcting for ISI) there was a main effect of confidence monitoring for Experiment 3.1 [$F(1, 34) = 20.87, p < .01$]. For Experiment 3.2, again confidence monitoring during took longer overall [$F(2, 56) = 24.60, p < .01$] as did source monitoring [$F(2, 56) = 39.01, p < .01$].

Cued recall performance. Error generation during learning enhanced target recall above both read conditions for all groups, as displayed in Figure 3.4. A 3 (learning condition: read short, read long, error generation) x 2 (source monitoring: yes, no) x 2 (confidence monitoring: yes, no) mixed ANOVA was run, with learning condition as the within-participants factor. There was an error generation effect, such that correct recall was greatest in the error generation condition (M = .74, SD = .17), above both read long (M = .56, SD = .20) and read short (M = .53, SD = .19) resulting in a main effect of learning condition [F (2, 112) = 64.42, MSE = .01, p < .001, η_p^2 = .54]. There were no significant interactions with confidence or source monitoring conditions [Fs <1].

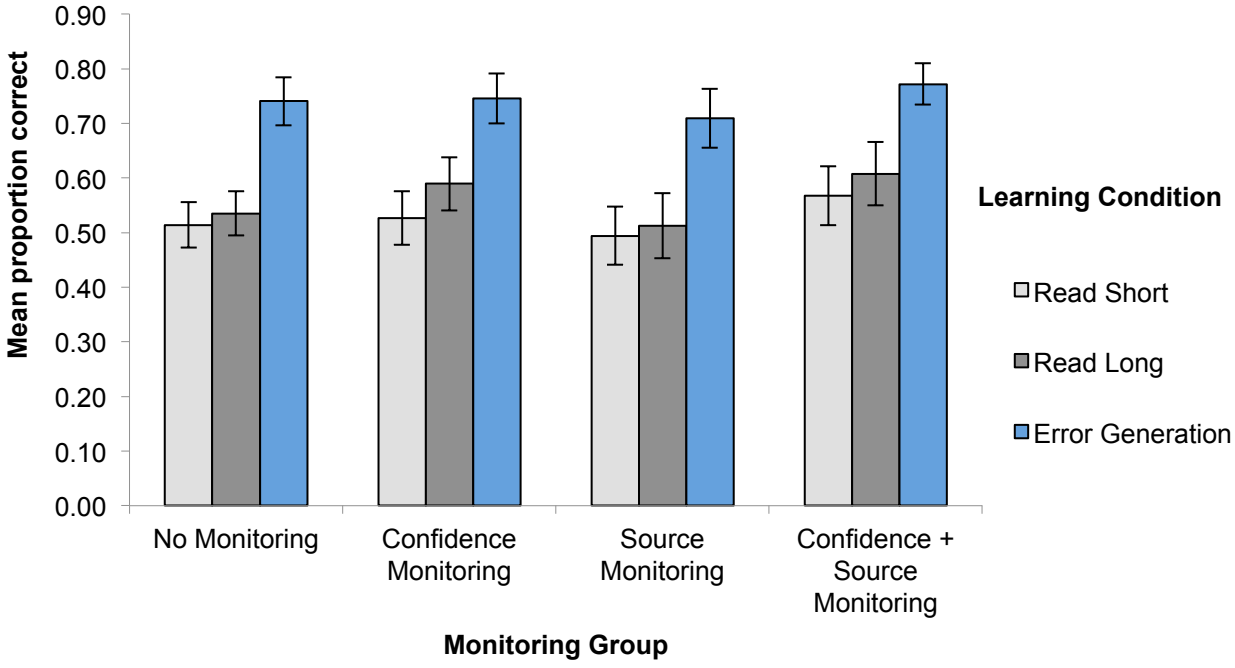


Figure 3.4 (Experiment 3b). **Cued Recall.** Mean cued recall performance for each of the four between participant groups. Performance beween groups was not altered as a function of confidence monitoring or source monitoring during cued recall. Overall, we still see robust error generation effects for all groups.

Item-by-item confidence judgments. Similarly to Experiment 3a, mean item-by-item confidence ratings were computed for participants who made item-by-item confidence ratings on the final cued recall test (confidence monitoring, and confidence + source monitoring groups), and are illustrated in Figure 3.5. (Cued recall performance is presented on the left panel and Item-by-item confidence score is presented on the right.) We ran a 3 (learning condition: read short, read long, error generation) x 2 (item type: cued recall, item-by-item confidence) x 2 (source monitoring: yes, no) mixed ANOVA to assess how accurate item-by-item confidence judgments were in correspondence to cued recall performance. Again, we found item-by-item confidence judgments on the final test were well calibrated with performance scores. Participants' item-by-item confidence judgments mirrored the demonstrated error generation effect found in cued recall performance, shown by a main effect of learning condition [F (1, 28) = 43.59, MSE = .02, $p < .001$, $\eta_p^2 = .61$] and overall, item-by-item confidence scores were similar to those of their actual correct performance, as there was no main effect of item type [F =1.29, p

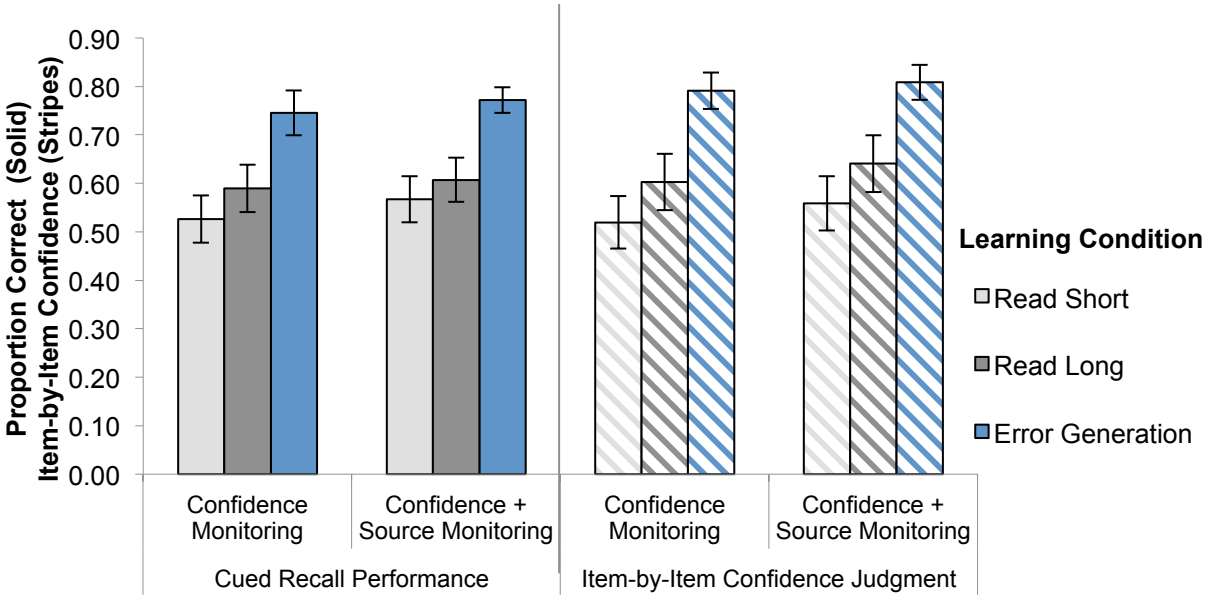


Figure 3.5 (Experiment 3b). **Cued Recall** (left panel, solid bars) **and Item-by-Item Confidence** (right panel, striped bars). Item-by-item confidence judgments were very accurate and showed similar patterns to cued recall for both monitoring groups who made item-level confidence judgments (confidence monitoring and confidence and source monitoring groups).

= .27]. Additionally, mean item-by-item confidence judgments did not differ as a function of making source judgments or not [F <1], nor did this interact with learning or item type [Fs<1]. [17]

Correlations between performance and item-by-item confidence. As described above, the means were similar for item-by-item confidence estimates compared to actual cued recall performance scores. Pearson correlations between individuals' mean performance and mean item-by-item confidence scores also confirmed that participants were well calibrated [$r_{\text{Read Short}}$ (30) =.83, $r_{\text{Read Long}}$ (30) =.87, $r_{\text{Error Generate}}$ (30) =.68, r_{overall} (30) =.85, ps< .001].

In addition, for each individual, a gamma was computed to assess if participants knew when they answered an item correctly or incorrectly. As a perfect relative metacognitive accuracy would be indicated by a gamma correlation of +1.0, participants were aware of which items they answered correctly or not, with an overall gamma of .93 (SD = .05). Mean gammas did not differ between the three learning conditions ($M_{\text{Read Short}}$ (30) =.91, SD = .10, $M_{\text{Read Long}}$ (30) =.93, SD = .10, $M_{\text{Error Generate}}$ (30) = .91, SD = .18) [F< 1].

Source judgment accuracy. Participants were capable of making source monitoring judgments identifying the learning condition for items presented at cued recall. They were quite accurate overall, correctly identifying the original learning condition on approximately 77% of the trials (SD = .18). Source memory for original learning condition was equally accurate between both those who monitored only source or both confidence and source [F < 1]. Critically, source memory was also similar across the three learning condition [F<1].[18] Read short and read

[17] However, there was a slight trend for participants to have higher item-level confidence judgments for error generation items than actual cued recall performance means [F (2, 28) =2.86, MSE = .01 p = .07, ηp^2 = .093] [error generation: $M_{\text{Confidence}}$ =. 80, SD = .12; $M_{\text{Cued Recall}}$ =.76, SD = .16 [t(30)=1.91, SE = .02, p = .07]. Note, this is in the opposite direction of underconfidence. For reading trials, there were no statistical differences between actual performance and mean item-by-item confidence judgments within learning conditions (read short: $M_{\text{Confidence}}$ = .54, SD = .21; $M_{\text{Cued Recall}}$ =.55, SD = .20, [t <1]; read long: $M_{\text{Confidence}}$ = .62, SD = .20; M_{Recall} =.63, SD = .20 [t <1]).

[18] We recognize this partially violates the rules of independence, however, as some participants indicated "don't know" these were no longer wholey dependent.

long trials were labeled correctly as "Together" on 75% (*SD* = .19) and 78% (*SD* = .21) of the time, respectively. In addition, participants correctly identified error generation items as having been studied in the "Separate" condition on 78% (*SD* = .23) of the trials. Therefore, overall, participants were quite good at identifying the source of specific trials.[19] In addition, making confidence judgments on the final test did not lead to differences in source accuracy [$F < 1$].

Global retrospective estimates of performance (GREPs). As before, participants reported the proportion of items they answered correctly in each of the three learning conditions. Each participant made these GREPs following the cued recall test. We previously found those who did not monitor performance during recall showed a metacognitive bias, and reported error generation to lead to worse recall than read long.

Therefore, we had two research questions to address in measuring GREPs for this current experiment. First, we wanted to replicate the findings of Experiment 3a: Was there a difference in GREPs for error generation as a function of any monitoring during recall versus not monitoring on an item-by-item? Secondly, were there differences in GREPs among these monitoring groups? To answer the first question, we collapsed over the three groups in which monitoring occurred during recall (confidence monitoring, source monitoring, confidence and source monitoring) and compared the mean GREPs of this collapsed group to the no monitoring group.

We again replicated the error generation underconfidence bias for the group who did not monitor during the test. Monitoring helped alleviate this underconfidence, and people in all three

[19] Conditions were re-named "Together Short", "Together Long" and "Separate." Though participants were quite accurate at identifying these as reading trials, they were not as accurate at identifying the timing length (long versus short). On the read short trials, participants labeled .45 as short (*SD* = .21) and .30 as long (*SD* = .19). On the read long trials, participants labeled .33 as long (*SD* = .18) and .30 as short (*SD* = .22).] Error generation trials were coded as correct if participants provided their original response, or if they wrote "S" or "separate" indicating they knew the original encoding task.

monitoring groups indicated that they had the highest recall in the error generation condition. Using a 2 (monitoring on cued recall: yes, collapsing over monitoring conditions, no monitoring) x 3 (learning condition: read short, read long, error generation) mixed ANOVA there was an interaction of monitoring and learning [F (1,58) = 6.53, MSE = .03, p = .02, η_p^2 = .10]. There were no differences between groups for GREPs of read short ($M_{\text{monitoring}}$ = .45, SD = .23; $M_{\text{no monitoring}}$ = .40, SD = .18) and read long ($M_{\text{monitoring}}$ = .55, SD = .23, $M_{\text{no monitoring}}$ = .50, SD = .20) (ts<1). However, those who monitored (either confidence, source, or both) estimated error generation to be better than those who did not ($M_{\text{monitoring}}$ = .63, SD = .24; $M_{\text{no monitoring}}$ = .40, SD = .26) [t (58) = 3.15, SE = .07, p = .003, d = .94]. For completeness we also report a significant main effect of learning condition, ($M_{\text{read long}}$ = .53, SD = .26; $M_{\text{error generation}}$ = .54, SD = .26; $M_{\text{read short}}$ = .43, SD =.24) [F (1,58) = 6.31, MSE = .03, p = .01, η_p^2 = .11] and a marginal effect of monitoring on mean GREPs, ($M_{\text{monitoring}}$ = .54, SD = .20, $M_{\text{no monitoring}}$ .44, SD = .20) [F (1, 58) = 3.50, MSE = .11, p = .07, η_p^2 = .06].

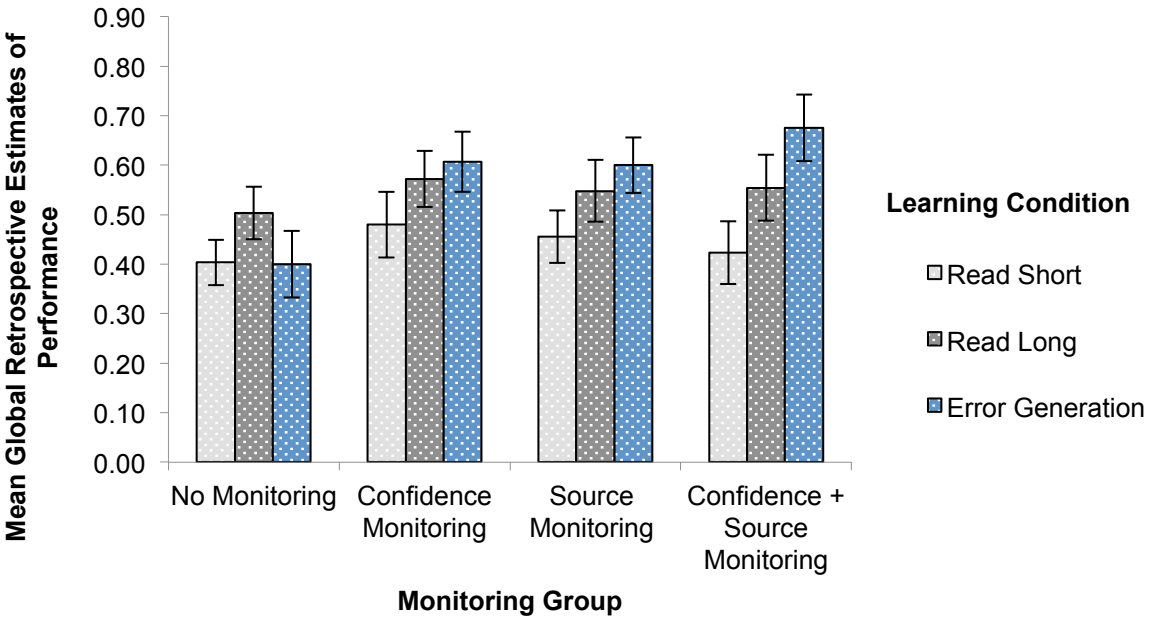


Figure 3.6 (Experiment 3b). **Global Metaocgnition.** Mean Global Retrospective Estimates of Performance (GREPS). When no monitoring occurred during the cued recall test, underconfidence in error generation is still evident. However, by confidence monitoring or source monitoring on the cued recall, participants' GREPs were higher than those who did not make item-level judgments.

Given that monitoring led to enhanced estimates of performance for the error generation strategy, our second aim was to assess potential differences in GREPs between the three unique monitoring conditions. We used a 3 (monitoring condition: confidence, source, confidence + source) x 3 (learning condition: read short, read long, error generation) mixed ANOVA. There was a significant main effect learning condition [F (2, 84) = 14.82, MSE = .02, p < .001, η_p^2 = .26] such that error generation (M =.63, SD = .24) GREPs were greater than read short (M =.45, SD = .23) [t (44) = 4.86, SE = .04, p < .001] and marginally higher than read long (M =.55, SD = .24) [t (44) = 1.96, SE = .04, p = .06]. Read long GREPs were also higher than read short [t (44) = 4.46, SE = .03, p < .001]. However, participants did not rate GREPs differently between these three monitoring conditions [no main effect of monitoring condition: F <1], nor was there an interaction with learning condition [F < 1].

Study strategy choice. Does item-level monitoring have consequences for future selection of study strategies? After making GREPs[20] participants were asked to state how many items (out of a possible 20) they would like to study in each of the three different learning conditions. Their aim was to allocate items across the three learning conditions (read short, read long, and error generation) in order to perform the best on a subsequent test (though this additional study-test session did not occur). For example, a participant might have indicated she wanted to study 10 items in read long, 5 in read short and 5 in error generation. Across participants, means for the number of items selected in each of the three learning conditions are presented for each of our between subject groups (no monitoring, confidence monitoring, source monitoring, confidence + source monitoring) in Figure 3.7

[20] And after explaining why they rated the learning conditions as they did (see Appendix for more details).

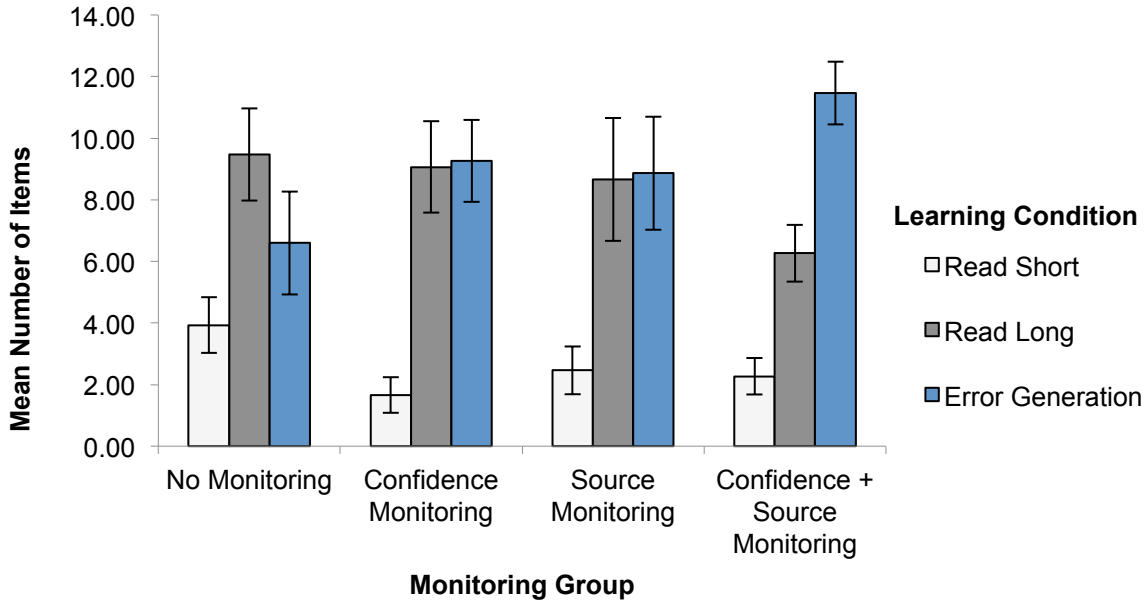


Figure 3.7 (Experiment 3b). **Strategy Selection.** Mean number of items selected for future study (out of a possible 20) for each of the three learning conditions, as a function of confidence and source monitoring during cued recall. Either confidence, source, or monitoring both during test led to a greater number of error generation items selected for an anticipated future study session.

For simplicity, our dependent measure is the mean number of items selected for study in the error generation condition. We followed similar analyses as with the GREPs, first analyzing if monitoring of any kind led to a greater number of error generation items being selected for a future study session compared to no monitoring at final test. Those who monitored selected marginally more items to study in the error generation condition ($M_{\text{monitoring}} = 9.78$, $SD = 5.57$) than those who did not ($M_{\text{no monitoring}} = 6.60$, $SD = 6.45$) [$t\,(58) = 1.87$, $SE = .06$, $p = .066$, $d = .55$]. In addition to the raw number items selected for the error generation condition, we also assessed the mean ranking across participants to see if those who monitored selected the most items for error generation relative to the other read conditions. In other words, what was the preferred condition for future study? Using a similar ranking procedure as in Huelser & Metcalfe (2012) we assigned the condition with the highest number of items for future study a score of 2,

the middle condition received a score 1, and the lowest a score of 0.[21] Using this relative rank method, we confirmed that those in the monitoring groups opted to select more error generation items (relative to the other two learning conditions) than those who did not monitor ($M_{\text{monitoring}}$ = 1.48, SD = .78, $M_{\text{no monitoring}}$ = .80, SD = 1.01) [Wilcox/Mann-WhitneyU: Z = 1.98, p = .048].

In addition to assessing if there was any effect of monitoring during recall on strategy selection, we wanted to determine if there were differences between the three monitoring groups (confidence monitoring, source monitoring, or both) on how many word pairs they would like to study in the error generation condition. Using a 3 (monitoring group: confidence monitoring, source monitoring, confidence + source monitoring) one-way ANOVA, we found no significant differences in how many items were allocated to future study in the error generation condition as a function of monitoring group (F<1). We also assessed relative rank of items selected for future study, and found no difference between these monitoring groups for selection of error generation items (Wilcox/Mann-WhitneyU, p = .39).

Discussion

Overall, in Experiment 3b we demonstrated that item-level monitoring during the criterion test enhanced global retrospective estimates for the effectiveness of the error generation condition, simply by making judgments of accuracy (confidence monitoring) or reflecting upon original learning strategy (source monitoring). As participants' accuracy of source was quite high and not different between the three learning conditions, we can assume overall that underconfidence in using errors as a learning strategy is not based on poor source monitoring of the learning condition. In addition to making confidence judgments, source monitoring can also

[21] For example, if a student assigned error generation 10 items for future study, read long 6 and read short only 4, error generation would be given a rank score of 2. We also accounted for ties by assigning half-points (e.g. If both a participant requested 8 items in each error generation and read long, and 4 in read short, the ranking scores would be 1.5, 1.5, 0 respectively.

help make a learner more aware of when error generation is beneficial. In fact, there were no differences among our three monitoring groups: those who monitored only confidence, only source, or did both monitoring tasks all showed higher global retrospective estimates of performance than those who did not monitor. In addition those who monitored during the criterion test tended to select more error generation items for a potential future study session, suggesting implications for metacognitive control.

General Discussion

In the current chapter, we demonstrated several novel findings on the metacognition of learning by making errors. In Chapter 1 (Huelser and Metcalfe, 2012), we demonstrated that participants did not rank error generation to be the best condition for learning. In Experiment 3a, we replicated these findings with global retrospective performance estimates, demonstrating that without monitoring at final test, learners estimated error generation to be less effective than reading the cue-target pairs for 10 seconds. However, when participants were asked to explicitly monitor their confidence in accuracy for each item on the criterion test, learners estimated error generation to be better than reading the items for a short period of time.

In Experiment 3b, we elaborated further on these findings, by demonstrating that participants were able to report source memory for how each item was originally learned. This supports that the underconfidence in learning by making errors was not due to an inability to judge source. Additionally, source monitoring also provided an opportunity for metacognitive knowledge updating, similar to confidence monitoring, as errorful learning was estimated to lead to enhanced retention above the control condition in which no monitoring took place. Still, we have not yet addressed why reflecting upon confidence and/or source might lead to decreased bias against error generation as an effective learning strategy.

Correctly Updating Metacognitive Knowledge

Although our methods of item-level monitoring can be used to update and correct metacognitive knowledge, the reason *why* monitoring during recall reduced the error generation metacognitive illusion is an important topic to explore. Veenman (2010) elaborated on the differences between metacognitive knowledge and metacognitive experiences. Global metacognitive knowledge is declarative-memory based information, subject to reconstruction and modification (Winne,1996; Winne & Hadwin, 1998; Winne & Jamieson-Noel, 2002), and is largely an inferential process (Herztog et al., 2008). Metacognitive experiences, on the other-hand, involve on-line monitoring of an experience during a task and are implicitly derived by a number of non-conscious sources (eg. liking, curiosity, surprise; see Efklides, 2006). Though metacognitive experiences might be implicit, once made consciously aware of the nature of these experiences, this declarative information then has the potential to be updated and become more stable metacognitive knowledge (Veenman, 2010), consistent with how other mental schema can be updated and revised (Winne & Jamieson-Noel, 2002). Given this line of reasoning, overtly responding to monitoring prompts forces participants to make these implicit item-level experiences of accuracy and source become explicit, highlighting the benefits of error generation. Thus, this monitoring allows one to correctly update metacognitive knowledge. Furthermore, in Hertzog et al.'s (2008) model based on encoding using imagery being more effective than rote study, several factors are outlined that are predictive of strategy knowledge updating. They highlight the importance of monitoring for updating metacognitive knowledge, and specifically monitoring on an *item-level at retrieval* to be contribute to overall global monitoring and strategy knowledge updating. Our findings on the reduction of the error

generation underconfidence bias are consistent with their model, and offer further support for the critical importance of monitoring during retrieval.

What is perhaps most interesting about the error generation underconfidence bias, and other recent work on overt-monitoring as a means of updating metacognitive knowledge, is that one might assume monitoring would occur naturally even without being prompted. The act of simply taking a test has been shown to be beneficial for updating metacognitive knowledge, in essence, because testing helps learners realize which items they are answering correctly or incorrectly (Koriat & Bjork, 2006; Hertzog et al. 2008). Even if item-level monitoring is taking place implicitly, making this process explicit by providing prompts could be "forcing" meta-level updating and evaluations of errors as an effective study strategy. However, difficult tasks beyond the ability of the learner will lead to a decrease in monitoring (Winne, 1996) and incorrect heuristics might be applied (Prins, Veenman & Elshut, 2006). It could be that our cued recall test itself is difficult[22], therefore, without explicit monitoring prompts, one might not be allocating additional cognitive resources to metacognitive monitoring and to updating metacognitive knowledge. This would be consistent with work that has shown enhanced calibration for higher performing students, (e.g. Bol & Hacker, 2001; Dunning, Kerri, Erlinger & Kruger, 2003) as perhaps they might tend to utilize more effective metacognitive strategies (Butler & Winne,1995; Winne, 1995, 1997). In addition, several studies suggest monitoring ability is of greater importance for metacognitive updating than intelligence (Pressley & Ghatala, 1990; Veenman, 2008; Schraw, 1994). Though beyond the scope of the current work, what might evolve is an interesting picture of the effects of performance and monitoring ability on the perception of the error generation benefit.

[22] It at least takes longer to monitor both confidence and source, as noted in Appendix B.

Implications for Metacognitive Control

Metcalfe (2009) summarized much of the work done on implications of metacognitive monitoring and how one's assessment of learning influences study strategy selection, suggesting a direct link between monitoring and metacognitive control. Determining if our manipulation of confidence and source monitoring at test also yielded improved strategy selection is of great importance. Data from Experiment 3b regarding the connection between item-level metacognitive monitoring and subsequent number of items selected for study in the error generation condition indicate that monitoring during test does seem to have implications for metacognitive control. Either monitoring how an item was encoded (source) or the level of accuracy (confidence) led to more error generation items allotted for future study than when no monitoring occurred on the final test. Metacognitive control is likely influenced by many other factors other than input from performance monitoring, such as beliefs about one's own abilities and task difficulty (Bandura, 1997; Dunlosky & Hertzog, 1998; Hertzog et al., 2008; Metcalfe, 2009; Pintrich, Wolters, & Baxter, 2000). Therefore, we believe our simple manipulations of item-level monitoring to be impressive; assessing performance and source led to enhanced metacognitive control.

In summary, the bias against learning by making errors is malleable. There are a number of possible research options to explore regarding the source of the error generation underconfidence effect, which we will briefly discuss in Chapter 4. While future research is needed to disentangle *the origin of* this metacognitive illusion, we demonstrated metacognitive knowledge about errors can be updated simply by monitoring performance and assessing how one originally learned information. This is a simple, yet impressive manipulation, to enhance awareness of when errors help learning. Prompts to monitor and assess performance were

sufficient to make a learner aware of the utility of learning by making errors, which otherwise, is severely underestimated in its effectiveness for learning.

Chapter 4

Conclusions and Future Directions

Conclusions and Future Directions

Research on the topic of learning by making errors has been conducted across varying fields: cognitive psychology, education, animal learning, clinical psychology, organizational behavior, and neuroscience. Yet, the question of when and why errors can help enhance learning is far from resolved. In part, this is a result of each field using its own methodology, materials, and situations (e.g. pubic or private; at school compared to the office). Furthermore, there are even differences in the definition of "error" (e.g. procedural, vs. declarative, severe or easily overcome), which makes a unified theory of error utility a difficult feat. If the goal is to be able to use this research to inform when learning is helpful and to understand why it is beneficial, dynamic collaboration is key.

The current body of work, which focuses on the foundations of the cognitive components of the error generation effect, is a critical first step in creating a groundwork upon which to build integrative theories about when and why error generation is beneficial for memory. We hope it inspires more research spanning multiple domains, as we are far from understanding the complicated and dynamic process of learning from errors. Many questions remain from both memory and metacognitive perspectives.

Error Generation and Remaining Questions

Memorial Mechanism

Throughout this dissertation, generating errors followed by correct answer feedback led to enhanced memory over simply studying the correct answer, even when given twice the amount of time to study the correct cue and answer pairing. As being able to recall one's original error led to better memory for the correct answer in Chapter 2, our results are consistent with episodic recollection as one role of generating errors plays in aiding retrieval. Yet, there are

many future directions to pursue to further elucidate the potential episodic role of the error. If retrieving the error affords episodic remembering of the original event, then we might expect enhanced context memory when the original error and/or correct answer is retrieved. This research is currently ongoing in the lab and draws a parallel to neuroimaging research. Showing the activation associated with a previous target is also present at the time of retrieval suggests activation of prior episodic memory for an unrelated (non-target) event (Kuhl, Bainbridge, & Chun, 2012). By adapting our current paradigms for imaging, these neurological data, in combination with behavioral data, would further inform the mechanism of the error generation effect, and might reveal interesting interactions between memory systems. More research is needed to fully understand the direct role of the error in these declarative tasks. It may even be the case that we cannot view the error's contribution as simply "episodic" or "semantic", as our understanding of memory systems is evolving and yielding a more complex and dynamic interplay between what was once viewed as dichotomous (Shohamy & Trurk-Browne, 2013; Shohamy & Wager, 2008). Perhaps in order to mentally time travel back to one's error and re-experience the errorful episode (Wheeler, Stuss, & Tulving, 1997), a meaningful context into which to bind this information is a pre-requisite (Conway, 2009). Furthermore, if episodic memory requires a sense of self and self-knowing (c.f. Wheeler et. al, 1997), there is still opportunity to elaborate on the role of the "self" in error generation. For example, if one remembers someone else's error, would the error still serve as a mediator? Pursuing this research would help inform theory, as well as, contribute to the practical implications of this research, such as when to implement an errorful strategy in the classroom.

Individual differences. In addition, further research of individual differences helps inform memory theory beyond the limited scope of errors. If episodic memory is critical for an

error generation benefit, further examining subgroups with episodic remembering deficits would help enrich our understanding of mechanism theory. For example, some recent work has shown mixed results for the effectiveness of an error generation strategy for learning (Middleton & Schwartz, 2012). However, the reason behind these results is still unclear. Is error generation ineffective as a study strategy for older adults because of source confusion and interference between the error and target (Anderson & Craik, 2006; for a review, see Spencer & Raz, 1995), or because those with episodic memory deficits can no longer recall their original error and original encoding episode (c.f. Dunlosky et al., 2008)? By examining various populations and the efficacy of error generation as a learning strategy, we can begin to more richly understand the mechanism of the error generation benefit.

Additional Analyses: Universality of the Error Generation Effect

In further thinking of potential individual differences, was error generation always best for everyone in our paradigm? While error generation led to higher rates of cued recall in nearly all the experiments presented, there is the possibility that when a participant claimed that "read long" was best, it *was* in fact the best condition for that individual. In order to answer this question, we examined the mean performance on the cued recall tests collapsed over Experiments 1a, 3a, and 3b. See Figure 4.1 for mean cued recall performance in each of the three learning conditions (read short, read long, error generation) as a function of which condition participants reported to be the "best" condition for learning (omitting participants who subjectively rated read short as best due to small sample size). These comparisons illustrate whether generating errors enhanced learning above simply studying for those who claimed read long to be the best condition for learning. Even for those who believed long to be best, error generation was a successful strategy.

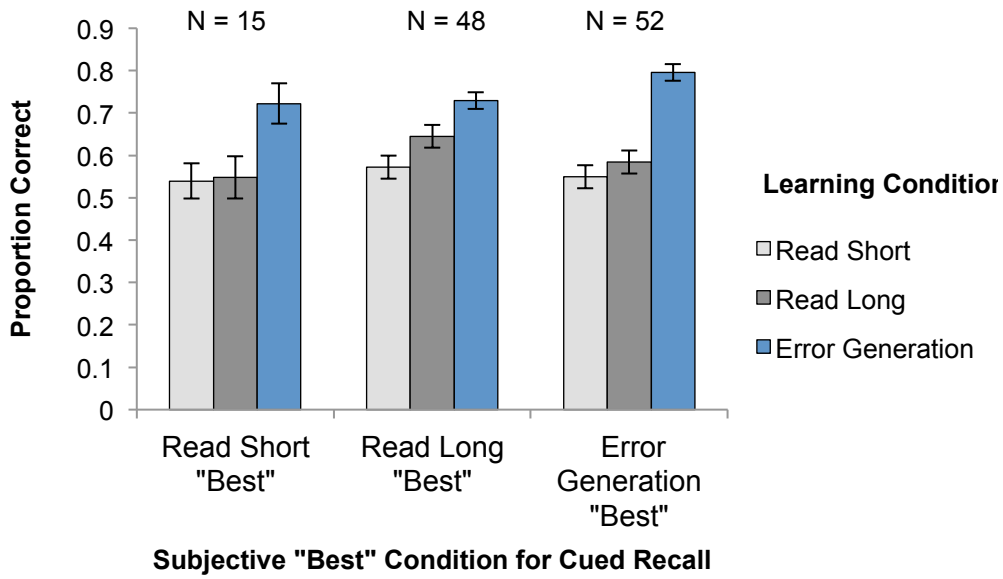


Figure 4.1 (Collapsed over Experiments 1a, 3a, & 3b). **Mean cued recall performance as a function of which condition was ranked/rated as "best" for learning.** Though error generation effect was largest when participants ranked as error generation best, the error generation effect is present regardless of which

We ran a 3 x (learning condition: read short, read long, error generation) x 2 (subjective preference: error generation best, read long best) mixed ANOVA on the proportion correct on the cued recall test. Overall, we see robust benefits of error generation on learning over read [F (2, 196) = 92.99, MSE = .01, p < .001, η_p^2 = .48]. Furthermore, it is important to note that those who rated error generation to be best for learning did not outperform those who rated read long to be best, as indicated by a lack of a main effect of subjective preference [F<1]. Yet, we found that even for those participants who "claimed read long was best", the error generation effect was still evident, but the benefit was larger for those who believed "error generation was best" (significant interaction [F (2,196) = 9.04, MSE = .01. p < .001, η_p^2 = .08]). Although our errorful learning task was beneficial for nearly everyone (only 17 participants actually performed better in read long out of 115 participants), these results also highlight that some learners benefit more

than others from efforful learning.[23] (See Appendix E for a scatter plot of the mean difference score of error generation above the mean cued recall of the read conditions.)

Along these lines, it is important to note the situation in in which these learners were asked to both study and take the test. Our current paradigm was rather innocuous (and, anecdotally, "fun" according to discussions with participants post-task). Participants were alone in a room, guessing paired associates, the data were anonymous, and the word "error" was never mentioned by the experimenter. It is hard to imagine a "lower risk" scenario. Therefore, it is perhaps not as surprising that error generation was beneficial, as hopefully there was minimal negative emotion to modulate performance. However, what if the experimenter never left the room and observed the errors? Or if participants were told the experiment was a test of intelligence? Our seemingly innocuous task suddenly might be interpreted differently, induce threat, and perhaps no longer be as universally beneficial (Chalabaev, Major, Sarrazin, & Cury, 2011). Understanding the factors that increase fit (Higgins, 2000) with error generation would likely have implications for both one's approach and ability to learn from errors (deLange & Kippenberg, 2009; Dweck & Legget, 1998). Furthermore, there is likely a dynamic interplay between cognitive abilities and emotion (e.g. emotion regulation, c.f. Ochsner & Gross, 2005) in determining when error generation will aid or hurt learning (see Zhao, 2011), and is of critical importance to explore.

Error Generation Underconfidence Bias

Still, we find it surprising that without monitoring performance, immediately post-task learners did not know that error generation was beneficial strategy for learning. In Chapter 3 we demonstrated that the error generation underconfidence bias does not exist on the item-level, but

[23] Note, overall performance between error generation as best and read long as best are not different from each other [$F = 1.12$, $p = .27$].

it instead appears when asked to reflect globally upon performance. Though we cannot currently disentangle the source of the underconfidence bias, we know it is malleable as we were able to reduce it through the item-level monitoring in Chapter 3.

Self-report data and questions about the source of the metacognitive bias. As a preliminary measure to seek further insight of the source of the error generation underconfidence bias, in Experiments 3a and 3b, we directly asked participants to report why a condition was best for learning. Summaries of these self-report data are presented in Appendices C and D, and highlight particular reasons that were included in the participants' responses. We anticipated that individuals might report, "I do not like making errors" as a plausible explanation for why the read long condition was best, but no participant overtly reported such a stark claim. This in part suggests that the bias might not be based simply upon a firmly-held pre-existing metacognitive belief that making errors is a negative.[24] Therefore, it is of interest to further pursue individual differences in the error generation underconfidence bias as a function of pre-experimental beliefs

Of note were participant reports that involved processes at encoding (e.g. more time and fewer distractions) for why read long was best. In essence, generating an error might seem more "effortful" compared to simply reading the word pairs during the encoding phase. Along similar lines, retrieval might not seem as fluent and straightforward for the error generation items, due to the possibility of the mental presence of the error at retrieval (also as suggested by the self-report data). Most interestingly, participants who selected the read long condition as best for learning described the error itself as interference, while participants who selected error generation as best

[24] Based on pilot data, it appears the students in the Columbia University participant pool do not have a negative bias against making errors, *in general*. See Appendix B for a list of means and standard deviations, t-test and p values for comparisons to a score of neutrality. Measures of Learning from Errors (e.g. "Mistakes help me to improve my work"), Competence (e.g. "When I have made a mistake, I know immediately how to correct it") and Error Risk Taking (e.g. "I'd prefer to err than to do nothing at all") are all rated above neutral (*p*s<.01), while Error Stress (e.g. "I find it stressful when I err") is neutral (t <1).]

described the error as a useful mediator, although both groups reported the mental presence of the error equally. Does this merely indicate a difference of interpretation by the participants, does it contribute to bias, and furthermore, might this suggest that individuals have different mechanisms for learning by making errors?

In addition to understanding why learners are originally underconfident with regards to the benefits of error generation, of great importance is the persistence of this metacognitive bias and the benefits of correctly updated metacognitive knowledge over an extended period of time. In other words, are there are long-term benefits of monitoring performance, and if so, does it make one more likely to select the most effective learning strategy in the future? The real world application of generating errors as a learning strategy may not be useful if item-level monitoring only updates metacognitive knowledge immediately, and a learner does not select to use the best strategy in the future. While we began to investigate this issue in Experiment 3b by asking participants to select which study conditions they would use on a future test, further experiments would help enrich our understanding of this critical question.

Empirical Questions to Inform Theory

Of course, one of the most apparent extensions of our current work is to vary the type of study materials used during experiments. Associative materials are useful in attempting to isolate and investigate boundary materials, but they are limited in ecological validity. Though some studies show promise of error generation effects using various materials (e.g. educational text (Hayes, Kornell & Bjork, 2009), various computer tasks (Keith & Frese, 2008, for a review), statistics (Swartz & Martin, 2004)) others have not been as successful (e.g. science materials, (McDaniel et al., 2010), trivia (Kang et al. 2011)). Are these mixed literature findings due to the task type (cued recall verses multiple choice), materials (general knowledge trivia, science text,

paired associates) or situation (during class, or at home, or online)? By investigating and isolating further boundary conditions, we can enrich our understanding of the error generation effect and its mechanisms, and begin to find real world application for this effective learning strategy.

Final Thoughts

We would like to end by restating an important caveat of this research: Errors made without receiving correct feedback are likely to remain incorrect (Butler, Karpicke & Roediger 2008; Fazio, Huelser, Johnson & Marsh, 2010; Metcalfe & Kornell, 2007; Pashler Cepeda Wixted & Roher, 2005; Pashler, Zarow & Triplett, 2003). Despite this limitation to learning through errors, the potential benefits of error generation for memory should not be overlooked. Despite the overwhelming post-task metacognitive illusion that errors are harmful to learning, this challenging strategy can lead to enhanced memory (consistent with Desirable Difficulties Framework (c.f. Schmidt & Bjork, 1992) and Region of Proximal Learning (Metcalfe & Kornell, 2002, 2003)). As one participant reported, "[Error Generation] was my best [option], because having to type out a word and then looking to see if my word was right ingrained the correct pain more in my memory." When you get corrective feedback, a little "pain" from making an error can enhance learning. Hopefully, simply by making basic metacognitive reflections, more learners will become aware that a little pain can go a long way.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, *24*, 1–3. doi:10.1016/j.learninstruc.2012.10.003

Anderson, N. D. & Craik, F.I.M. (2006) The mnemonic mechanisms of errorless learning. *Neuropsychologia,* 44, 2806-2813. doi:10.1016/j.bbr.2011.03.031.

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(3)*, 261-285.

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

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

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.

Arnold, K. M., & McDermontt, K.B. (2013). Test-potentiated learning: distinguishing between

direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(3)*, 940-945.

Barnes, J. M, & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology, 58,* 97-105. ratings. *Journal of Verbal Learning & Verbal Behavior, 19,* 338-368.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 553-563*.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31,* 297-305.

Benjamin, A. S. & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), Implicit Memory and Metacognition (pp. 309–338). Mahwah, NJ: Erlbaum.

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

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.

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

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

Bjork, R. a, Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–44. doi:10.1146/annurev-psych-113011-

143823

Butler, A. C., Karpicke, J. D., & 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. (2001). The correction of errors committed with high confidence. *Metacognition and Learning, 1*, 1556-1623.

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

Craik, F. I. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & 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, 104*, 268-294.

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 contribute to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(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. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. Journal of Experimental Psychology: Learning, Memory, & Cognition.

Chalabaev, A., Major, B., Sarrazin, P., & Cury, F. (2012). When avoiding failure improves performance: Stereotype threat and the impact of performance goals. *Motivation & Emotion,* 36, 130–142. *doi:*10.1007/s11031-011-9241-x

Chase, C.C. (2012). The interplay of chance and skill: Exploiting a common game mechanic to enhance learning and persistence. *Proceedings of the 2012 International Conference of the Learning Sciences.*

Clare, L., & Jones, R.S. (2008). Errorless Learning in the Rehabilitation of Memory Impairment: A Critical Review. *Neuropsychology Review, 18(1)*, 1-23.

Collins, A.M., & Loftus, E.F. (1975). A spreading activation of semantic processing. *Psychology Review, 82(6)*, 401-428.

Collins, A.M., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L.W. Gregg (Ed.), *Cognition in Learning and Memory* (117-138). New York: John Wiley.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628. doi:10.1016/j.jml.2005.08.005

Conway, M. A. (2009). Episodic memories. *Neuropsychologia*, *47*(11), 2305–13. doi:10.1016/j.neuropsychologia.2009.02.003

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 & Verbal Behavior, 7,* 613 -616.

de Lange, M. A., & Van Knippenberg, A. (2009). To err is human: How regulatory focus and action orientation predict performance following errors. *Journal of Experimental Social Psychology*, *45*(6), 1192–1199. doi:10.1016/j.jesp.2009.07.009

Daw, N.D. & Shohamy, D.(2008). The cognitive neuroscience of motivation and learning. *Social Cognition, Special Issue: Cognitive Motivation and Motivated Cognition,* 26, 593-620.

Dunlosky, J., Baker, J. M. C., Rawson, K. A., & Hertzog, C. (2006). Does aging influence people's metacomprehension? Effects of processing ease on judgments of text learning. *Psychology and Aging, 21,* 390 – 400.

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The Contribution of Mediator-Based Deficiencies to Age Differences in Associative Learning. *Developmental Psychology, 41(2)*, 389-400.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Beverly Hills, CA: SAGE.

Dunlosky, J. & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding student's judgments can improve their achievement. *Learning and Instruction*, *24*, 58-61. doi:10.1016/j.learninstruc.2012.05.002

Dunlosky, J., Serra, M., & Baker, J. M. C. (2007). *Metamemory*. In F. Durso, R. Nickerson,S. Dumais, S. Lewandowsky, & T. Perfect (Eds.), Handbook of Applied Cognition (2nd ed.,pp. 137–159). New York, NY: Wiley.

Dweck, C. S. (2006). Mindset: The new psychology of success. New York: Random House.

Dweck C. S, & Leggett, E. L. (1998). A social-cognitive approach to motivation and personality. *Psychological Review, 95,* 256–273.

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

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making, 18,* 199-212.

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 and Review*, *16,* 88-92.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory & Language*, 58, 19-34.

Fischhoff, B. & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*, 155-172.

Flavel, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906-911.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

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.

Green, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. Review of Educational Research. doi:10.3102/003465430303953.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74,* 1464-1480.

Greenwald, A.G., Nosek, B.A., (2009). Attitudinal dissociation: what does it mean? In: Petty, R.E., Fazio, R.H., Briñol, P. (Eds.), *Attitudes: Insights from the New Implicit Measures*. Erlbaum, Hillsdale, pp. 65–82.

Grimaldi, P.J., & Karpicke, J.D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40(4),* 505-513.

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

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. a. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. doi:10.1037//0022-0663.92.1.160

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: a focus on calibration. In J. Dunlosky, & R. Bjork (Eds.), Handbook of memory and metacognition. Mahwah, NJ: Lawrence Erlbaum Associates.

Hajcak, G., & Foti, D. (2008). Errors are aversive: Defensive motivation and the error-related negativity. *Psychological Science*, *19*(2), 103-108.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning & Verbal Behavior, 6*, 685-691.

Hays, M.J., Kornell, N., & Bjork, R.A. (2013). When and why a failed test potentiates the

effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(1)*, 290-296.

Hertzog, C., Kramer, A. F., Wilson, R. S., and Lindenberger, U. (2008). Enrichment effects on adult cognitive development: Can the functional capacity of older adults be preserved and enhanced? *Psychological Science Public Interest, 9*, 1–65.

Hertzog, C., Price, J., & Dunlosky, J. (2008). How is knowledge generated about memory encoding strategy effectiveness? *Learning and Individual Differences*, *18*, 430-455.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. Kruglanski, (Eds.), *Social psychology: Handbook of basic principles* (pp.133– 168). New York: Guilford Press.

Higgins, E.T. (2000). Making a good decision: Value from fit. *AmericanPsychologist, 55*, 1217–1230.

Higham, P. A. (2013). Regulating accuracy on university tests with the plurality option. *Learning and Instruction*, *24*, 26–36. doi:10.1016/j.learninstruc.2012.08.001

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

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: repetition and recursive reminding. *Memory & Cognition*, *32*(2), 336–50.

Hintzman, D. L. (2011). Research Strategy in the Study of Memory: Fads, Fallacies, and the Search for the "Coordinates of Truth." *Perspectives on Psychological Science*, *6*(3), 253–271. doi:10.1177/1745691611406924

Huelser, B.J. & Metcalfe, J. (2010, November). When does initial retrieval failure lead to later

success? *Poster presented at the 51st annual meeting of the Psychonomic Society, St. Louis, MO.*

Huelser, B.J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40(4)*, 514-527.

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.

Jacoby, L.L., & Wahlheim, C. N. (2013). On the Importance of Looking Back: The Role of Recursive Remindings in Recency Judgments and Cued Recall. *Memory & Cognition, 41(5)*, 625-637.

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

Kahana, M. J., Howard, M. W., & Polyn, S. M. (2008). Associative retrieval processes in episodic memory. In H. L. Roediger III (Ed.), *Learning and memory: A comprehensive reference:* Vol. 2. Cognitive Psychology of Memory. Oxford: Elsevier.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *The American Psychologist, 58*(9), 697–720. doi:10.1037/0003-066X.58.9.697

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.

Keith, N. & Frese, M. (2008). Effectiveness of error management training: a meta-analysis. *The Journal of Applied Psychology*, *93*(1), 59-59. doi: 10.1037/0021-9010.93.1.59

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory & Language*, 32, 1-24.

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., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*, 959-972.

Koriat, A., Lichtenstein, S., Fischhoff, B. (1980). Reasons for Confidence. Journal of Experimental Psychology: Human Learning and Memory 6(2), 107-118.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. Journal of Experimental Psychology: General, 131(2), 147-162.

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., & Metcalfe, J. (2014). The effects of memory retrieval, errors and feedback on learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.). *Applying science of learning in education: Infusing psychological science into the curriculum* (225-

251). Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/asle2014/index.php

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449-468.

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.

Kuhl, B. A, & Chun, M. M. (2012). Attending to the present when remembering the past. *Neuron*, *75*(6), 944–7. doi:10.1016/j.neuron.2012.09.002

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.

Lichtestein, S. and Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance. 26,* 149-171.

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

Manis, Shelder, Jonides, & Nelson (1993). Availability heuristic in judgments of set size and frequency of occurrence. *Journal of Personality & Social Psychology*, *65* (3) (1993), pp. 448–457.

Marcel, T. (1980). Recognition of Polysemus Words: Locating the Selective Effects of Prior Verbal Context. In R.S. Nickerson (Ed.), *Attention and Performance VIII* (435-457). New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Matvey, G., Dunlosky, J., Shaw, R. J., Parks, C., & Hertzog, C. (2002). Age-related equivalence and deficit in knowledge updating of cue effectiveness. *Psychology & Aging*, 17,589-597.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1263-1274.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (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.

McCloskey, M. & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General, 114*, 1-16.

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

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. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131,* 349-363.

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 and Review, 15*,174-179.

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

Metcalfe, J. & Kornell, N. (2003). The Dynamics of Learning and Allocation of Study Time to a Region of Proximal Learning. *Journal of Experimental Psychology: General, 132*, 530-542.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52,* 463–477.

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

Metcalfe, J., & Mischel, W. (1999). A hot/cool system analysis of delay of gratification: Dynamics of willpower. *Psychological Review,106* (1), 3 –19.

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: A critical review. *Neuropsychological Rehabilitation, 22*(2), 138-168. PMID:22247957

Miele, D. B., Molden, D. C., & Gardner, W. L. (2009). Motivated comprehension regulation: Vigilant versus eager metacognitive control. *Memory & Cognition, 37*, 779-795.

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 Academic Press.

Neely, J.H. (1976). Semantic priming and retrieval from lexical memory: evidence for facilitatory and inhibitory processes. *Memory & Cognition, 4(5)*, 648-654.

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

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin,95,* 109-133.

Nelson, T. O. (1999). Cognition versus metacognition. In R. J. Sternberg (Ed.), The nature of cognition (pp. 625–641). Cambridge: MIT.

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, 2(4),* 267–270.

Nelson, T.O. & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. *The Psychology of Learning and Motivation*, *26,* 125-173.

Nietfeld, J., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal of Experimental Education, 74*(1), 7–28.

Nietfeld ,J.L.,Cao, L.,and Osborne, J. W.(2006).The effect of distributed monitoring exercises

and feedback on performance, monitoring accuracy,and self-efficacy. *Metacognition and Learning* 1, 159–179.

Ochsner, K. N. & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences, 9(5),*242-249.

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

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.

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

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

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A Context Maintenance and Retrieval Model of Organizational Processes in Free Recall. *Psychological Review, 116*(1), 129–156. doi:10.1037/a0014420

Posner, M.L., & Snyder, C.R. (1975). Facilitation and inhibition in the processing of signals. In

P.M.A. Rabbitt & S. Domic (Eds.), *Attention and Performance*. New York: Academic Press.

Pyc, M. A., & Rawson, K. A. (2009). Testing the Retrieval Effort Hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437-447.

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

Ratcliff, R., & McKoon, G. (1994). Spreading activation versus compound cues theory of memory. *Psychology Review*, *101(1)*, 185-187.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing. of Experimental Psychology. Learning, Memory, and Cognition, 28(1), 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.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology*, *82*, 416-425.

Rybowiak, V., Garst, H., Frese, M., & Batinic, B. (1999). Error Orientation Questionnaire (EOQ): reliability, validity, and different language equivalence. *Journal of Organizational Behavior*, *20*, 527–547. Retrieved from

http://phdnet.unigiessen.de/wps/pgn/dl/showfile/ebme_de/1144/Error_orientation_questionnaire.pdf

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology*: *General*, *134*, 124-128. doi:10.1037/0096-3445.134.1.124

Schneider, W. & Pressley, M. (1997). Memory development between two and twenty. Mahwah, NJ: Erlbaum.

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.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psycology, 19*(2), 143-154. doi: 10.1016/j.bbr.2011.03.031.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, *7*(4), 351–371. doi:10.1007/BF02212307

Schraw, G., & Nietfeld, J. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology*, *90*(2), 236–248. doi:10.1037//0022-0663.90.2.236

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.

Schulz, R. A. (1996). Focus on form in the foreign language classroom: Students' and teachers' views on error correction and the role of grammar. *Foreign Language Annals, 29*, 343-364.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review, 20,* 463–467.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and

feeling of knowing. *Psychonomic Bulletin & Review*, **1**, 357-375.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for learning: The hidden efficiency of original student production in statistics instruction. *Cognition & Instruction, 22*, 129-184.

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14,* 332–348.

Shohamy, D. & Turk-Browne, N (2013) M*echanisms for widespread hippocampal involvement in cognition*., *Journal of Experimental Psychology: General, 142(4),* 1159-1170.

Skinner, B. F. (1968). Technology of teaching. Engelwood Cliffs, NJ: Prentice Hall.

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.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory & Language*, *26*, 589-607.

Spencer, W. D., & Raz, N. (1995). Differential Effects of Aging on Memory for Content and Context : A Meta-Analysis, *10*(4), 527–539.

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. (2005). Metacognitive control: Children's short-term versus long-term study strategies. *Journal of General Psychology*, *132*, 347-363.

Son, L. K. (2010). Metacognitive control and the spacing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 36*(1), 255-262.

Son, L. K. & Kornell, N. (2008). Research on the allocation of study time: Key studies from1890

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.

Son, L. K., & Schwartz, B. L. (2002). *The relation between metacognitive monitoring and control*. In T. J. Perfect, & B. L. Schwartz (Eds.), Applied metacognition (pp. 15–38). Cambridge: Cambridge University Press.

Son, L. K., & Sethi, R. (2006). Metacognitive control and optimal learning. *Cognitive Science*, *30*, 759-774.

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.

Sparrow, B. and Wegner, D.M. (2006). Unpriming: The deactivation of thoughts through expression. *Journal of Personality and Social Psychology, 91 (6),* 1009-1019

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

Paas, F., & Sweller, J. (2011). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*(1), 27–45. doi:10.1007/s10648-011-9179-2

Terrace, H. S. (1963a). Discrimination learning with and without "errors". *Journal of the Experimental Analysis of Behavior, 6(1*), 1–27.

Terrace, H. S. (1963b). Errorless transfer of a discrimination across two continua. *Journal of the Experimental Analysis of Behavior, 6,* 223–232.

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.

Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to student mistakes. *Teaching and Teacher Education*, *33*, 56–68. doi:10.1016/j.tate.2013.02.003

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–42. doi:10.3758/s13421-012-0274-5

Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1130.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and psychology of choice. *Science*, *211*, 453-458.

Van Hout-Wolters, B. (2000). Assessing active self-directed learning. In R. Simons, J. van der Linden, & T. Duffy (Eds.), *New learning* (pp. 83–101). Dordrecht: Kluwer.

Veenman, M. V. J. (2010). Learning to self-monitor and self-regulate. In R.E. Mayer & P.A. Alexander (Eds.) *Handbook of Research on Learning*, (pp. 197–218). New York: Routledge.

Veenman, M. V. J., Van Hout-Wolters, B. A. M., & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. doi:10.1007/s11409-006-6893-0

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109. doi:10.1016/j.learninstruc.2003.10.004

Veksler, V. D., Grintsvayg, A., Lindsey, R., & Gray, W. D. (2007). A proxy for all your

semantic needs. *Proc CogSci 2007* (pp. 2702-2702). Retrieved from: http://cwl-projects.cogsci.rpi.edu/msr/.

Wahlheim, C.N., & Jacoby, L.L. (2013). Remembering Change: The Critical Role of Recursive Remindings in Proactive Effects of Memory. *Memory & Cognition*, *41(1)*, 1-15.

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

Wheeler, M. a, Stuss, D. T., & Tulving, E. (1997). Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychological Bulletin*, *121*(3), 331–54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9136640

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: Lawrence Erlbaum Associates, Inc.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, *8*(4), 327–353. doi:10.1016/S1041-6080(96)90022-9

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), Metacognition in educational theory and practice (pp. 277-304). Hillsdale, NJ: Lawrence Erlbaum.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27(4),* 551–572.

Zimmerman, B. (2000). Self-Efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*(1), 82–91. doi:10.1006/ceps.1999.1016

Zhao, N. B. (2011). Learning from errors : The role of context, emotion, and personality. *Journal of Organizational Behavior, 463*, 435–463. doi:10.1002/job

Appendix A

Error Attitudes (Columbia University, Spring 2014). Fifty-one Columbia University Students' Error Attitudes. The first four items are mean compontents from Error Orientation Questionnairre (EOQ) [Rybowiak et al (1999)]. Ratings have a neutral score of 0 (ranging from -2 to +2). A postive value indicates agreement, while negative indicates disagreement.

		M	*SD*	*t*	*p*
	Error Orientation Questionnaire Constructs:				
	EOQ: Error Competence, e.g. "know how to correct"	0.58	0.75	5.55	0
	EOQ: Learning from Errors, e.g. "help me improve"	0.75	1.27	4.21	0
	EOQ: Error Risk Taking, e.g. "prefer to err than do nothing"	0.75	0.86	6.24	0
	EOQ: Stress from Errors , e.g. "stressful when I err"	0.02	1.03	0.14	0.89
-	Making errors on an assignment helps me figure out what I need to study for the test.	0.99	1.03	6.85	0
-	When I'm confident that I know the correct answer, making a mistake is particularly disappointing.	0.91	1.17	5.55	0
-	I tend to remember questions that I previously made mistakes on.	0.9	1.02	6.29	0
-	A mistake is really just an opportunity for learning.	0.65	1.34	3.45	0
-	Making a mistake often leads me to the correct answer.	0.59	1.18	3.55	0
-	Putting effort into recalling an answer is helpful even if it leads to a mistake.	0.55	1.17	3.35	0
-	Making a mistake often helps when learning how to perform a new skill.	0.52	1.28	2.89	0.01
-	Making a mistake before I learn the correct answer sometimes improves my learning.	0.37	1.2	2.23	0.03
-	I'm careful to avoid making mistakes whenever I'm learning something.	0.29	1.19	1.77	0.08
-	When something is complicated, I'm not particularly bothered by making mistakes.	0.26	1.21	1.56	0.13
-	When approaching a new activity, I like to jump in and learn from my mistakes.	0.25	1.18	1.54	0.13
-	Making mistakes motivates me to continue learning.	0.25	1.26	1.39	0.17
-	When I get everything right, it means I'm not challenging myself enough.	-0.12	1.35	0.62	0.54
-	Trial and error is my preferred learning strategy.	-0.19	1.29	1.03	0.31
-	Having my mistakes corrected feels embarrassing.	-0.2	1.4	1	0.32
-	Making a mistake is only helpful if I find out the correct answer right away.	-0.36	1.34	1.94	0.06
-	When I make a lot of mistakes, I feel like giving up.	-0.44	1.13	2.8	0.01
-	Making an incorrect guess initially can get in the way of remembering the correct answer later.	-0.47	1.25	2.68	0.01
-	When I make a mistake, I become distracted and unable to focus on the task at hand.	-0.54	1.13	3.42	0
-	It does not bother me to make mistakes	-0.56	1.13	3.55	0
-	When I try to do something on my own before learning the correct answer, I only become more frustrated.	-0.76	1.08	5.04	0
-	I find it difficult to remember the correct answer after having made a mistake.	-0.87	1	6.24	0
-	A mistake represents a personal shortcoming.	-1.19	1.2	7.04	0

Appendix B

Mean Reaction Time Data

Reaction times for Experiments 3a and 3b as a function of Trial Type (Cued Recall, Confidenece, or Source Judgments). Total Trial time accounts for ISI.

	Trial Type During Criterion Test							
	Cued Recall		Confidence Judgment		Source Judgment		Mean Total Trial (correcting for ISI)	
Condition	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Experiment 3.1								
Control (No Monitor)	4010.77	1216.3	.	.	.	.	4760.77	1216.3
Confidence Ratings	5120.11	1205.1	1768.22	500.8	.	.	6888.33	1557
Experiment 3.2								
Control (No Monitor)	4305.25	928.08	.	.	.	.	5305.25	928.08
Confidence Ratings	5042.43	1152.1	1600.14	385.35	.	.	7142.57	1429.6
Source Monitoring	5158.13	1286.6	.	.	1986.45	467.88	7644.6	1552
Conf+Source	4949.14	1030.7	2090.53	583.32	2452.92	653.64	9676.89	1940.9

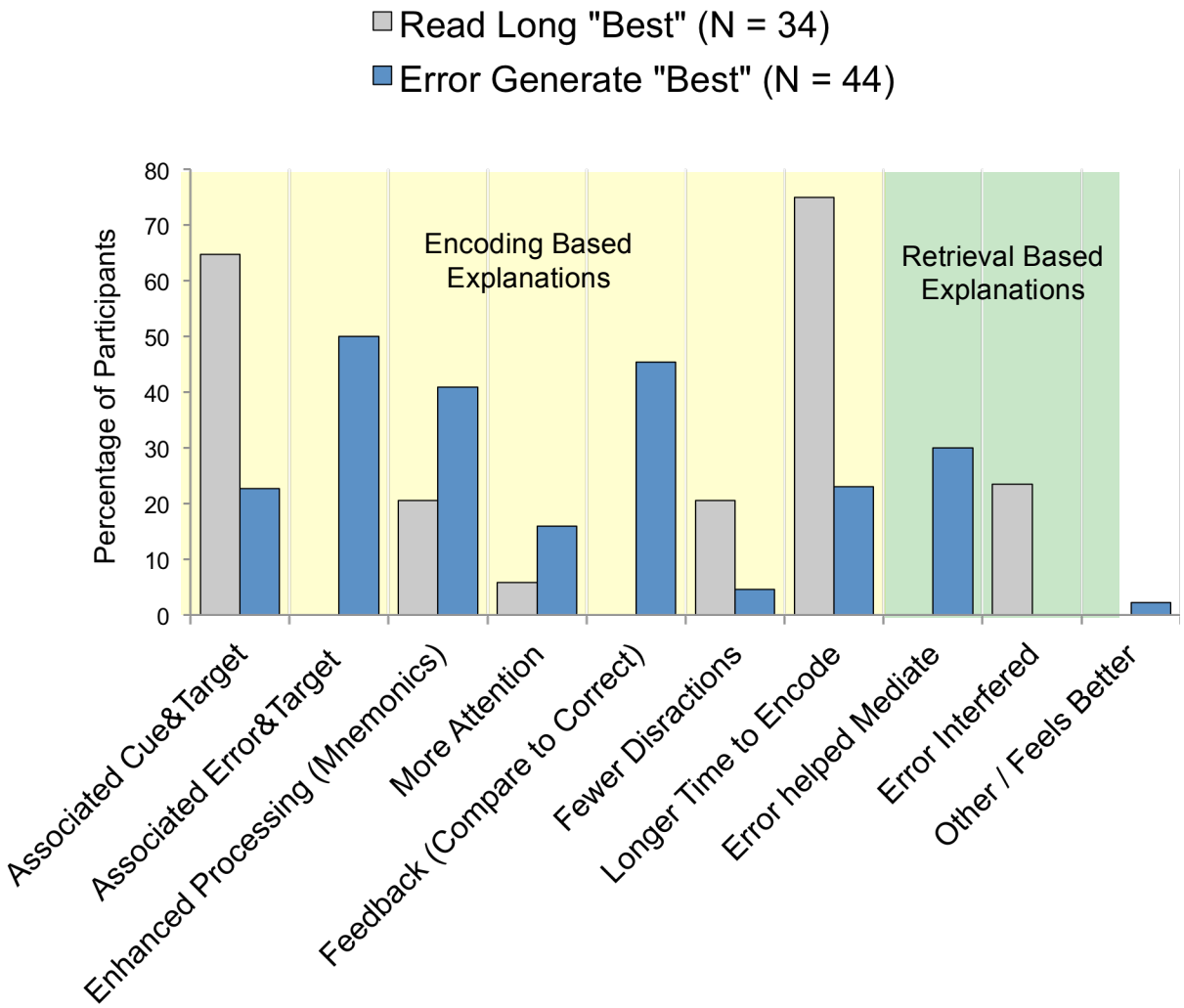
Critically, there were no differences in RTs to produce the target across the between subjects monitoring conditions for both Experiments [Fs<1]. However, on total trial time (even when correcting for ISI) there was a main effect of confidence monitoring for Experiment 3.1 [$F(1, 34) = 20.87$, $p < .01$]. For Experiment 3.2, again confidence monitoring during took longer overall [$F(2, 56) = 24.60$, $p < .01$] as did source monitoring [$F(2, 56) = 39.01$, $p < .01$].

Appendix C

Favored Learning Condition Self Report Data (Experiments 3a and 3b)

Percentage of participants who reported each of the strategies above to explain why their selected strategy was best for learning. Sample responses for each category are listed in Appendix D. Note, the total will sum to greater than 100% as participants reported a mean of 2.28 reasons (*M* = 2.28, *SD* = .75).

Read short was excluded, as only 11 participants reported this as the best learning strategy and our primary interest was between read long and error generation.



Appendix D

Favored Learning Condition Self Report Data :

Sample Free Report Responses and ChiSquare Values

Chi Squares (χ^2) between error generation or read long reported as best for condition, within each category. Note, despite differences in valence of the error, both groups reported similar rates of the error at retrieval (30% of error generation vs. 24% of read long) [$\chi^2 = 1.03$, $p = .31$].

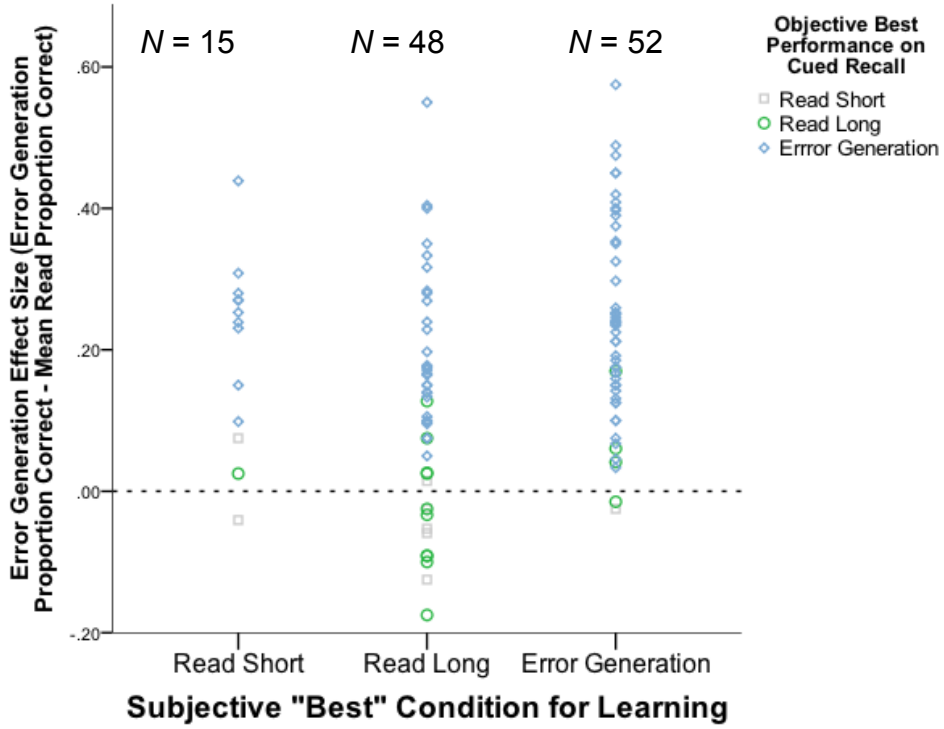
Category	Sample Free Responses (Self-Report)	χ^2
Other / Feels Better	For some reason, it was a lot easier to recall the thought processes that occurred	0.78
Error Interfered	If I guess wrong, I might remember the wrong answer instead of the correct answer	11.35 **
Error helped Mediate	I generally remembered what I had said; I knew why I got that wrong and then I remembered it going further for the last part	14.35 ***
Longer Time to Encode	The longer display of words lead to the best performance because I had more time to learn the pair; More Exposure	17.86***
Fewer Distractions (or more for other)	Distracted me from learning the correct word; If something distracted me, I risked barely or not making the connection at all	4.84*
Feedback (Compare to Correct)	Thinking about how this word was different from the one you typed in; Entered a word different from the eventual second word helped me to remember the second word	20.78***
More Attention	Kept the mind occupied; Paid more attention	1.89
Enhanced Processing (Mnemonics)	I have to think about the word and its meaning and engage with it; Less about rote memory, and more about forming the precursors of a literal relationship between the two	3.64*
Associated Error&Target	I associated my answer and the correct answer	23.67***
Associated Cue&Target	Absorb both words; Process meaning of the word pair	13.96***

*$p < .05$, ** $p < .01$, *** $p < .001$
Self-report data were categorized by an independent coder who was not aware of the hypotheses of the experiment. Categories were created based on participant responses. up to three reasons were coded for each participant (M=2.16, SD = .84), therefore the proportion will add to greater than 1.00.

Appendix E

Scatter Plot Data of Error Generation Effect Benefit Sixe

(Mean cued recall performance of error generation – Mean cue recall of reading conditions)



Error Generation Benefit Size

Difference score of cued recall performance in the error generation condition above the mean cued recall of the read conditions [Error generation Performance – (Mean (Read Long, Read Short) Performance]. Collapsing across Experiments 1a, 3a, and 3b (all weakly related word pairs). Data are parsed into bins according to which condition a participant reported as being subjectively "best" for performance. Though semi-redundant with the y-axis, the data points are colored by which condition was actually best for learning. Any data point above 0 indicates that error generation led to a benefit in performance above mean cued recall collapsed over read conditions. It is of interest to note that the metacognitive illusion is uni-directional: Rarely do participants report read long as best when error generation was best for cued recall performance.