DISSERTATION

# STRATEGIC ENCODING AND EPISODIC DISCRIMINATION (SEED) MODEL OF ERROR CORRECTION

Submitted by

Hannah Hausman

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2020

Doctoral Committee:

Advisor: Matthew G. Rhodes

Anne Cleary
Dan Graham
James Folkstead

# ABSTRACT

# STRATEGIC ENCODING AND EPISODIC DISCRIMINATION (SEED) MODEL OF ERROR CORRECTION

Despite what many students and teachers believe, making errors while learning can improve long-term learning of correct information. This paper proposes the Strategic Encoding and Episodic Discrimination (SEED) model of error correction, which proposes that in comparison to errorless learning, making errors while learning enables individuals to effectively adapt how they encode the correct answer and then, on a later memory test, use episodic memory to discriminate between the correct answer and other information that may be retrieved. Experiment 1 tested the strategic encoding component of SEED and found that errorful learning enhanced memory relative to errorless learning, but the benefits of errorful learning could not be explained by strategic adaptations in study times. Experiment 2 tested both the strategic encoding and episodic discrimination components of SEED and contrasted SEED with other accounts of error correction. The results of Experiment 2 were largely consistent with SEED and revealed that errorful learning enhanced memory by both increasing the likelihood that the correct answer was retrieved on the final test and improving participants' ability to distinguish between correct and incorrect answers.

# ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Matt Rhodes, for not only his guidance and help with this project, but for his mentorship and support throughout my time in graduate school.

# TABLE OF CONTENTS

# CHAPTER 1 – INTRODUCTION

*"A person who has never made a mistake never tried anything new."*

- Albert Einstein

Making mistakes is inevitable when trying something new, whether it is learning how to juggle, how to speak a new language, or how to write a persuasive speech. An important practical question is how making mistakes affects long-term learning. If the goal is to eventually perfect a skill, increase knowledge, or deepen understanding, should initial learning and practice conditions be designed to invite or avoid mistakes? For example, if your goal is be fluent in Norwegian, should you avoid making mistakes by looking up every word and its pronunciation before, say, ordering food at a restaurant? Alternatively, will you ultimately improve your knowledge of the language if you practice ordering food at a restaurant from memory, even though you will inevitably make some vocabulary or pronunciation mistakes as you learn?

Students and teachers constantly make decisions about whether to invite or avoid errors during learning, albeit often implicitly. For example, students can decide whether to try to answer difficult questions in class or wait until the teacher presents the answer. Students can try to solve challenging math problems on their own or copy answers from the internet or their peers. When studying for a test, students could make flashcards to test themselves, which may elicit errors; alternatively, students could read the bold terms in the textbook and avoid producing incorrect information. Indeed, most students wait to test themselves while studying until they know the information well, suggesting students believe errors should be avoided while learning (e.g., Hartwig & Dunlosky, 2012; Janes, Dunlosky, & Rawson, 2018; Karpicke, 2009;

Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2007; Morehead, Rhodes, & DeLozier, 2016).

Similarly, when delivering a lecture, teachers can accurately explain a concept immediately or have students try to generate their own explanations first but invite the possibility that students' explanations will be inaccurate. Intensive observational studies of teaching strategies have revealed cross-cultural differences in how errors are handled in K-12 math classes (e.g., Stevenson & Stigler, 1994). In the United States, teachers primarily teach correct procedures and spend little time addressing students' mistaken thinking. In contrast, teachers in Japan intentionally invite errors through strategically designed questions and discuss mistakes in students reasoning as a key learning tool.

Broadly speaking, this paper addresses the question of whether errors should be invited or avoided during learning in order to prevent mistakes in the future when performance matters, say, on an exam. Specifically, the present studies test a new theoretical explanation for how making errors affects learning of correct information.

## The Influence of Errors on Learning: Typical Methods and Results

Early behaviorist accounts of learning suggested that errors should be avoided during practice because producing mistakes would only further reinforce those mistakes in memory, thereby increasing the chances that the mistakes would be repeated in the future when accurate performance matters (Ausubel, 1968; Bandura, 1986). For example, if a student incorrectly guessed that Sydney is the capital of Australia while studying, the hypothesis was that they would become more likely to repeat that Sydney is the capital of Australia on their geography test. Despite initial concerns that making mistakes would entrench errors and impair subsequent performance, ample evidence has suggested that errors should not be avoided during practice, but rather, should be invited as a potent learning opportunity. Metcafe (2016) concluded in a comprehensive review of the literature that "errors can greatly facilitate new learning."

The general paradigm for studying the role of errors in learning uses variants of the following conditions (Figure 1). Participants are given an opportunity to answer a question and then, after some delay, are provided with correct answer feedback. On a later test, the question is presented again and participants are asked to recall the correct answer. In studies that rely on conditional analyses, memory for the correct answer on the test is compared between questions for which an error was initially generated and questions for which the correct answer was initially retrieved.

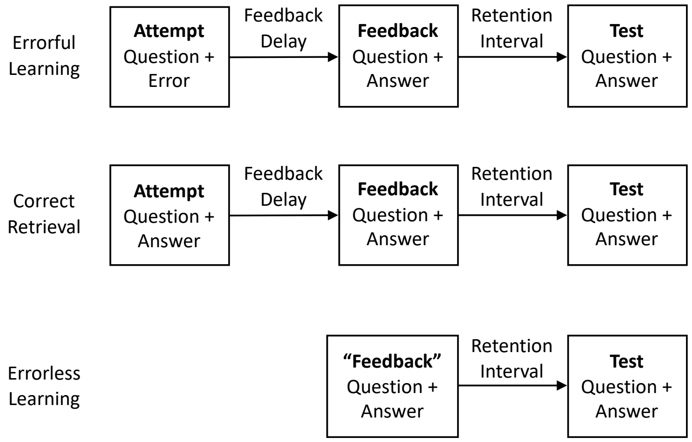


*Figure 1.* Typical paradigm for studying the effects of errors on learning.

For example, Pashler, Zarow, and Triplett (2013) had participants learn the definitions of obscure vocabulary words (e.g., cygnet, declivity). After learning the definitions, participants made an initial attempt to remember the vocabulary; the definition was presented along with the first two letters of the vocabulary word. For some definitions, participants were able to generate the correct vocabulary word (correct retrieval) and for other definitions participants were not able to generate the correct vocabulary word (errorful learning), at which point the correct vocabulary word was presented. One week later, participants took a final test on the vocabulary words. When participants made an error on the initial practice test, they were also more likely to make an error on the final test in comparison to when they correctly retrieved the vocabulary term on the initial test. However, one cannot conclude that producing an error *caused* poorer long-term learning because of item-selection effects. That is, the questions that participants answered incorrectly initially may have simply been more difficult for the participant, leading to lower final test performance as well. Pashler and colleagues concluded that, "To know what causal impact an error had, uncontaminated by item selection issues, one would need to compare later performance after the subject makes an error on an item with performance on other items for which an error would have been made— but for which no test was even given" (p. 1056).

Therefore, another typical and preferable (Kornell, Hays, Bjork, 2009; Pashler et al., 2003) paradigm involves randomly assigning questions (or participants) to an errorful or errorless learning condition. In the errorless learning condition, there is no initial attempt to generate an answer and only the correct answer is presented (Figure 1). In this paradigm, researchers typically use questions that participants do not know the answers to so that every attempt results in producing an error. For example, Kornell and colleagues (2009) had participants learn fictional, yet plausible trivial questions (e.g., What fabled cat sprang to new life from its grave? Lynx). These trivia questions probed fictional knowledge (e.g., there is no such fabled cat), but they sounded like they could be real. In the errorless learning condition, participants only studied the fictional questions and the "correct" answers. In the errorful learning condition, participants were shown the question and guessed the answer before the "correct answer" was presented. Because the questions were fictional, participants' guesses were necessarily errors. On a later test, the question was presented and participants were asked to recall the correct answer. Producing a mistake initially enhanced subsequent learning of the correct answer compared to errorless learning. Participants remembered the answers to 31% of the fictional trivia answers if they only read the question and answer but remembered 41% of the fictional trivia answers if they had made an incorrect guess first. Thus, research that experimentally manipulates error production during initial learning has revealed that making mistakes facilitates subsequent learning of correct information (for a review, see Metcalfe, 2017).

Yet ample evidence has also revealed that sometimes errors have no effect on, and sometimes they can even impair subsequent learning (Bridger & Mecklinger, 2014; Cyr & Anderson, 2015; Grimaldi & Karpicke, 2012; Hausman & Rhodes, 2018; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Vaughn & Rawson, 2012). For example, Grimaldi and Karpicke (2012) had participants learn related word pairs (e.g., tide-wave). In the errorless condition, participants only studied the correct pair (e.g., tide-wave). In the errorful condition, participants first made a guess for what the second word could be (e.g., tide-wa_____; many participants guessed water) before studying the correct pair. Trials were omitted if participants guessed the correct answer. On a later test, participants were shown the first word (e.g., tide) and were asked to recall the correct second word (e.g., wave). Participants were significantly more likely to recall the correct answer on the later test in the errorless condition than in the errorful condition.

Thus, making a mistake does not universally improve subsequent learning of the correct answer. Even when participants initially learn the correct answer after making a mistake, participants' original mistakes can return over time (Butler, Fazio, & Marsh, 2011; Metcalfe &

Miele, 2014). Any theoretical explanation of how errors are corrected must account for the mixed evidence in the literature as to whether initially producing an error enhances, has no effect on, or impairs learning of correct information. To preview, this dissertation develops a new theory of error correction (Strategic Encoding and Episodic Differentiation; SEED), which can account for the varying results of making errors that have been previously observed. The experiments test SEED alongside the most popular current account of error correction (errors-as-mediators).

## Errors as Mediators: An Existing Account of Learning from Errors

There is one particularly common explanation for why making an error would enhance subsequent learning of correct information. The errors-as-mediators account proposes that errors act as mediators, or steppingstones, providing an additional retrieval route from the question to the answer (Figure 2a; Bridger & Mecklinger, 2014; Cyr & Anderson, 2015, 2018; Kornell et al., 2009; Hays et al., 2013; Knight et al., 2012; Kornell, 2014; Vaughn & Rawson, 2012). For example, if an individual is asked, "What is the capital of Australia?," they may incorrectly guess Sydney before learning that the capital is Canberra. According to the errors-as-mediators hypothesis, on a later test, the individual can remember Canberra either directly from the question or first remember Sydney, which will subsequently activate the answer Canberra. In contrast, errorless learning does not enhance memory to the same degree because it only strengthens the question→answer association and the learner does not form an additional retrieval route from the question to the answer (Figure 2b).

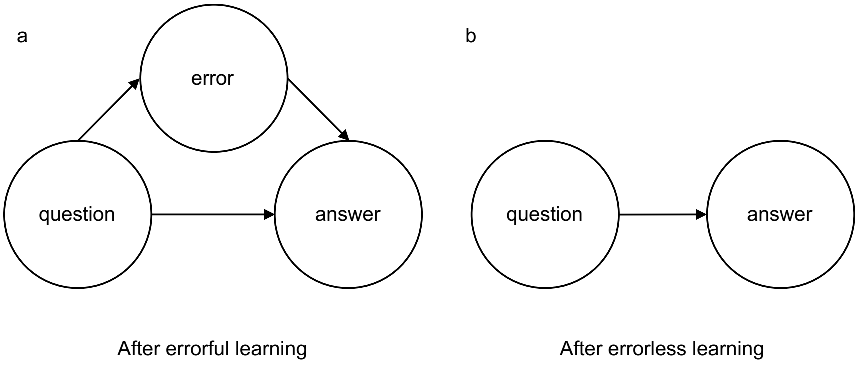


*Figure 2.* Proposed associations in memory among the question, error, and correct answer after errorful learning (a) and errorless learning (b), according to the errors-as-mediators hypothesis. Errorful learning is thought to enhance memory for the answer due to the creation of an additional retrieval route from the question to the answer: a mediated pathway from the question to the answer via the error that was made. Errorless learning is thought to be less effective because it results in only one way to retrieve the correct answer: directly from the question.

The errors-as-mediators hypothesis is a semantic account of how errors are beneficial for learning. That is, the errors-as-mediators account explains how making a mistake enhances learning of the correct answer by forming meaning-based associations among the cue, error, and answer in semantic memory. On a subsequent test, when the question is presented, it is thought that activation spreads through the network of related ideas and concepts (e.g., Collins & Quillian, 1972) directly to the answer and indirectly by activating the error, which activates the correct answer in turn (e.g., Kornell et al., 2009). In its current formulation, the errors-as-mediators hypothesis does not invoke ideas about episodic memory—i.e., consciously recalling the initial study phase in which the error was made and the correct answer was presented—as a way to learn from errors.

### ***Evidence for the Errors-as-Mediators Hypothesis***

The primary form of evidence for the errors-as-mediators hypothesis comes from manipulations of the semantic richness or meaningfulness of the learning materials, consistent with its foundation as a semantic account of learning from mistakes. The logic is that errorful conditions should not be as beneficial for long-term learning when the error and correct answer cannot be easily associated because the complete question→error→answer retrieval route is less likely to form. For example, consider a participant presented with the cue *door* who guesses *lock*. If the correct answer is *exit*, then *lock* and *exit* can be easily associated because the cue, error, and target are all semantically related. On a later test, the mediated route *door→lock→exit* can facilitate retrieval of *exit*. In contrast, if the correct answer is *shoe*, it is more difficult to associate with the guess *lock* and the additional retrieval route *door→lock→shoe* is less likely to form and cannot support later retrieval of *shoe*.

Indeed, errorful learning, compared to errorless learning, has been shown to enhance memory for related but not unrelated word pairs (e.g., Cyr & Anderson, 2015; Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012; Kornell, 2014). For example, Knight and colleagues (2012) had participants learn related (e.g., door-exit) and unrelated (e.g., door-shoe) word pairs under errorless or errorful conditions. In the errorless condition, participants only studied the correct pairs. In the errorful condition, participants guessed the target word based on the cue (e.g., door-???) before being presented the correct answer. Participants' guesses were nearly always errors and trials were excluded if the participant correctly guessed the target. On a subsequent final test, the relative effects of errorless versus errorful learning depended upon the type of material. Errorful learning enhanced memory for the correct answer for related word pairs, but errorless learning enhanced memory for the correct answer for unrelated word pairs.

Knight and colleagues (2012) also reported more direct evidence for the errors-as-mediators hypothesis. On the final test, participants were asked to recall both their original guess and the correct answer. Among the test trials on which participants could recall their original guess, participants were more likely to also recall the correct answer for related word pairs, but not unrelated word pairs. This conditional analysis suggests that error→answer association was more likely to form or was stronger for related pairs than unrelated pairs. One explanation is that it is difficult to form a complete cue→error→target mediated retrieval route for unrelated pairs because participants' guesses and the targets were not semantically related. Indeed, the benefits of errorful learning (e.g., band-???) may be larger when the participant's guess (e.g., drum) is more closely semantically related to the target (e.g., guitar vs. rubber; Cyr & Anderson, 2018; but see Metcalfe & Huelser, 2020).

### ***Limitations of the Errors-as-Mediators Hypothesis***

Although there is consistent evidence that manipulating the semantic relatedness of materials moderates the benefits of errorful learning, there is little causal evidence that it results from the ability to form the error→answer association. Manipulating the materials does not necessarily imply mediation was manipulated. The semantic relatedness of one's error and the correct answer may merely be epiphenomenal to memory for the correct answer (for a similar argument, see Karpicke, Lehman, & Aue, 2014). Some other feature of pair relatedness may affect one's ability to learn from errors besides the degree to which forming a mediated cue→error→answer path is afforded. However, any alternative theoretical account of learning from errors must account for the finding that producing errors enhances subsequent learning of related but not unrelated information.

There is also evidence that is difficult to explain in terms of the errors-as-mediators hypothesis. If making errors enhances learning by creating an additional retrieval route, then generating multiple errors should be more beneficial than generating one error. Prior research does not support this prediction (Lehman & Karpicke, 2016; Vaughn & Rawson, 2012). For example, Vaughn and Rawson (2012) had participants learn related word pairs (e.g., athletic-muscle) under errorless or errorful conditions. In the errorful conditions, participants generated a guess (e.g., athletic-???) either once or three times before studying the correct pair. On a subsequent cued-recall test, memory for the correct target was better in the errorful than the errorless condition regardless of the number of guesses that had been generated initially. Thus, generating more errors, and presumable more retrieval routes, did not enhance memory.

Another prediction of the errors-as-mediators hypothesis is that making one's error more memorable on the final test should improve memory for the correct answer because it increases the likelihood that the mediated retrieval route can be used to recall the correct answer. However, experimentally manipulating memory for the error on a subsequent test also does not reliably improve the benefits of errorful learning (Vaughn & Rawson, 2012). In short, empirical evidence is mixed as to whether errors enhance learning by serving as a mediator from the question to the answer on retrieval attempts and much of the data supporting the errors-as-mediators hypothesis is also correlational in nature.

In addition to empirical limitations of the errors-as-mediators hypothesis, the errors-as-mediators account also raises a theoretical conundrum. Memory is best when a retrieval cue uniquely specifies the target. The more information associated with cue, the lower the likelihood the target will be retrieved (Moscovtich & Craik, 1976; Nairne, 2002; Raaijmakers, 2003; Raaijmakers & Shiffrin, 1981; Watkins & Watkins, 1975). The errors-as-mediators account is therefore inconsistent with the cue-dependent nature of memory. Errors-as-mediators suggests that memory for the correct answer is better because the cue becomes associated with *more* information—i.e., the error *and* the answer—not less information (for a similar critique, see Grimaldi & Karpicke, 2012).

In light of the empirical and theoretical limitations of the errors-as-mediators hypothesis, a new theory for how individuals learn from making mistakes is warranted. Any such theory should invoke the idea that producing an error enhances subsequent memory of the correct answer by some mechanism that results in the cue more uniquely specifying the target (Figure 3a). In contrast, errorless learning could lead to poorer learning if the cue becomes associated with multiple ideas, including the correct answer (Figure 3b).

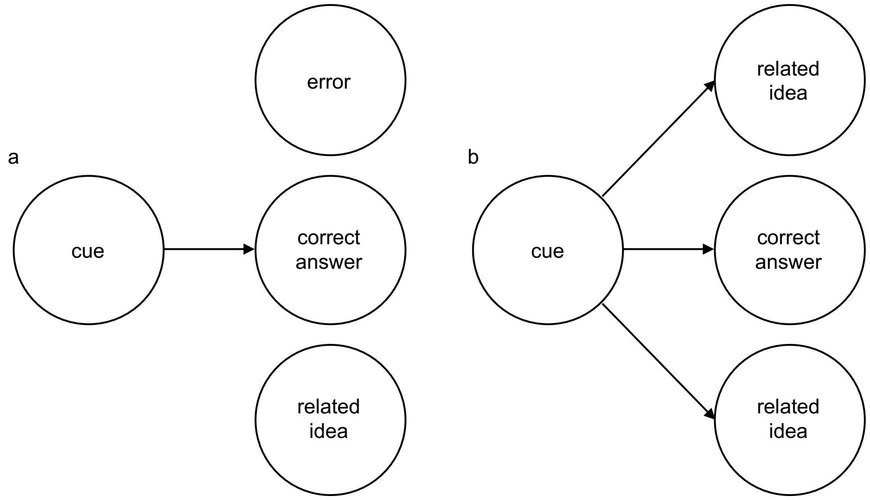


*Figure 3*. The foundational assumptions of a theory of the benefits of learning from errors that comports with cue-uniqueness. Errorful learning should lead to a more unique association between the cue and the correct answer at the exclusion of one's original error and related ideas (a). Errorless learning should lead to the cue becoming associated with multiple related ideas, including the correct answer (b).

Accordingly, I propose a theory of learning from errors that is consistent with the principle of cue-uniqueness, which can also account for the effects of pair relatedness on learning from errors. This theory will be tested against the errors-as-mediators hypothesis.

## Strategic Encoding and Episodic Differentiation Model

I propose the *strategic encoding and episodic differentiation* (SEED) model as an alternative to the errors-as-mediators account of how making errors enhances learning. SEED suggests that there are two routes by which making an error can improve memory for the correct answer: strategic encoding and episodic differentiation. The strategic encoding and episodic differentiation components will be explained in detail in the following sections. To preview, strategic encoding refers to the idea that, relative to errorless learning, making a mistake encourages learners to better encode the correct answer by effectively adapting how they study the answer once it is presented. Examples of strategic encoding include increasing attention, spending more time studying, or changing the nature of the information that is encoded (e.g., generating a mental image rather than repeating the information over and over again). Learning the correct answer better initially would improve memory for the answer on subsequent tests.

However, on a subsequent test, multiple ideas and possible answers may come to mind and the learner must select the correct answer from among them. Episodic discrimination refers to the idea that, relative to errorless learning, making an error initially may increase episodic memory for the study experience, or make it "stand out" in memory. Anecdotally, we have all experienced a situation in which we cannot forget the time we made an embarrassing mistake. More formally, making mistakes while learning may enable learners to later consciously recall the event in which they made a mistake and then studied the correct answer. Therefore, when learners have multiple possible answers in mind, the would be able to use episodic memory to recall which one had been presented in the experiment as the correct answer.

Thus, SEED suggests that making errors can enhance memory for the correct answer by enabling individuals to 1) learn the correct answer well initially and 2) identify the correct answer among possible answers that come to mind on subsequent tests. Experiment 1 tested the strategic encoding component of SEED and Experiment 2 will test both the strategic encoding and episodic differentiation components of SEED. These two routes to accurate memory for the correct answer closely align with broader theories of how information is recalled, namely, the two-stage model of retrieval.

### *Two-Stage Model of Retrieval*

SEED is based on the well-developed two-stage model of retrieval (e.g., Jacoby, Kelley, & McElree, 1999), which proposes front-end and back-end retrieval processes. Broadly speaking, the front-end processes involve generating possible answers. Back-end retrieval processes involve evaluating the accuracy of the possible answers and selecting the correct answer from among them.

**Front-End Retrieval Processes.** When a question or cue is presented and memory is searched for the answer, front-end retrieval processes dictate what information comes to mind during that memory search. Accurate memory performance depends upon the correct answer being included in the set of information produced by the memory. Therefore, one front-end retrieval strategy is to "cast a wide net," or bring to mind as much relevant information as possible to increase the likelihood that the correct answer will be produced. Many theories of memory are consistent with the idea that retrieval will cast a wide net. Specifically, associative models of memory suggest that searching memory for one piece of information would activate semantically related information in memory (e.g., Anderson, 1996; Collins & Quillian, 1972; Raaijmakers, 2003). For example, trying to retrieve the capital of Australia may also activate the names of other cities in Australia and around the world.

The errors-as-mediators account proposes that casting a wide net on a test is critical for retrieving the correct answer and benefiting from having made an error. In order for the cue→error→answer mediated pathway to facilitate retrieval of the correct answer on a test, the error must be included in the search set so that it can activate the answer in turn.

However, casting a wide net may not always be an effective front-end retrieval strategy. Merely producing the correct answer will not lead to accurate memory test performance if the correct answer cannot be distinguished from other candidate answers in the search set. Therefore, an effective front-end retrieval strategy might be to narrow the scope of one's memory search to include only the correct answer and exclude non-target information.

Different memory phenomena suggest front-end retrieval processes exist that restrict a memory search to only the target information. For example, competing non-target information can be suppressed (i.e.., actively made less accessible in memory) in order to retrieve only the target information (e.g., Anderson, Bjork, & Bjork, 1994, 2000).  In addition, Jacoby and colleagues (Halamish, Goldsmith, & Jacoby, 2012; Jacoby, Kelley, & McElree, 1999; Jacoby, Shimizu, Daniels, & Rhodes, 2005; Jacoby, Shimizu, Velanova, & Rhodes, 2005) have demonstrated how individuals are able to engage in a front-end retrieval process referred to as *source-constrained retrieval* in order to narrow the memory search to only include information from a particular source (e.g., one list of studied items but not another). Jacoby and colleagues proposed that individuals are able to qualitatively change how they engage in retrieval in order to constrain the search set to relevant information.

For example, Jacoby and colleagues (2005) had participants study two lists of words. For one list, participants rated how pleasant the meaning of each word was, a "deep" encoding strategy known to enhance memory. For another list, participants indicated whether the word contained an O or a U, a "shallow" encoding strategy know to have little benefit for memory. Next, participants completed two recognition tests. On the "deep" recognition test, participants were shown the words from both the pleasantness and vowel lists as well as never-before-seen words, known as foils. Participants were instructed to indicate whether each word had been studied in the pleasantness list. Similarly, on the “shallow” recognition test, participants were shown the words from both the pleasantness and vowel lists as new foils. Participants were instructed to indicate whether each word had been studied in the vowels list.

The researchers were not interested in the effects of deep or shallow encoding on the memory for the initial lists of words, per se. Rather, the researchers were investigating the retrieval process involved in the recognition test. Therefore, the experiment ended with a surprise “memory for foils” recognition test. On the memory for foils test, participants were presented with the foils from the deep recognition test and the shallow recognition test, as well as new never-before-seen words. Participants indicated whether they had seen each word at any point in the experiment.

Recognition was better for foils that had been presented during the deep recognition test than foils that had been presented during the shallow recognition test. This pattern of results is surprising because the foils were presented in the same manner and thus should have been encoded in the same way on both deep and shallow memory tests. Intuitively, participants should have read the foil, not recognized it as a previously-studied word, and rejected it.

Contrary to this intuition, the results suggest that on the deep recognition test, participants initiated front-end retrieval processes. That is, participants recapitulated deep processing (i.e., thinking about pleasantness) as a way to search memory but constrain it to only words in the pleasantness list so that they could determine whether each word on the recognition test was in that search set. As a result, participants inadvertently engaged in deep processing of the foils on the deep recognition test, producing better memory for these foils on the critical, memory-for-foils test. In contrast, on the shallow recognition test, participants likely recapitulated shallowing processing (i.e., thinking about vowels) to generate a search set that contained only words from the vowels list. In this case, participants likely inadvertently engaged in shallow processing of the foils on the shallow recognition test, producing worse memory for these foils down the line.

Although the memory-for-foils paradigm is complex, it provides convincing evidence that individuals qualitatively alter how they engage in retrieval in order to bring to mind relevant information and exclude irrelevant information for the retrieval task at hand. Thus, Jacoby and colleagues' work shows that one route to accurate memory is through front-end retrieval processes that narrow the search set, or constrain it to only target information.

**Back-End Retrieval Processes.** Back-end retrieval processes refer to processes by which one selects the correct answer among the candidate answers that are produced as a result of front-end retrieval processes. The two-stage model of retrieval suggests that individuals produce a candidate answer and then evaluate the accuracy of that candidate. If the individual judges the candidate answer to be correct, the memory search will stop; otherwise, the memory search will continue (e.g., Halamish, Jacoby, Goldsmith, 2012; Koriat & Goldsmith, 1996; Koriat, Goldsmith, & Halamish, 2008). Thus, the two-stage model of retrieval proposes an iterative process of generating a candidate answer (a front-end retrieval process) and evaluating its accuracy (a back-end retrieval process) until one believes a correct answer has been produced or that continuing to search memory would no longer be fruitful (e.g., Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Nelson & Narens, 1990), leading the search to be terminated (another back-end retrieval process).

The back-end processes of evaluating the accuracy of candidate answers and deciding whether to continue or terminate the memory search are examples of *metacognition* (i.e., thinking about one's own thinking; e.g., Flavell, 1979). Metacognition consists of two key components: *monitoring* and *control*. Monitoring refers to the process of self-assessing one's learning, understanding, or skill; control refers to the self-regulated behaviors made as a result of monitoring (Dunlosky, Mueller, & Thiede, 2016; Nelson, 1996; Nelson & Narens, 1990; Rhodes, 2019). Evaluating the accuracy of a candidate answer involves metacognitive monitoring and is therefore often referred to as *post-retrieval monitoring*. Although it has not been explicitly stated in previous research, the errors-as-mediators hypothesis implies that accurate post-retrieval monitoring is essential to benefiting from errors. If the cue activates the error, which activates the correct answer in turn, individuals must be able to monitor which piece of information is the error and which is the correct answer in order to report the correct answer on a test.

One product of monitoring is a decision to continue or terminate the memory search, invoking metacognitive control that is often referred to as *post-retrieval control*. Another example of a post-retrieval control process is the decision of whether to report or withhold an answer once the memory search has stopped and the best candidate answer has been selected. Koriat and Goldmsith (1996) suggested that individuals monitor the accuracy of retrieved information and decide what information to report in order to maximize the accuracy of their memory reports. For instance, they offer the practical example of instructions in a courtroom for a person testifying to “tell the truth, the whole truth, and nothing but the truth.” In the context of providing testimony, a large number of details may come to mind. In deciding which details to report in the testimony, the individual must weigh the risks of providing incorrect information with not providing complete testimony. Post-retrieval monitoring involves assessing the likelihood that each of the details is accurate. Post-retrieval control involves deciding which of those pieces of information to report in testimony and which pieces of information to withhold based on their likelihood of being correct.

Evidence for the existence of post-retrieval monitoring and control as back-end retrieval processes comes from experiments in which participants are asked to report their confidence in their retrieved information and/or make decisions about which information to provide because it is likely correct and which information to withhold because it is likely incorrect (e.g., Halamish et al., 2012; Kelley & Sahakyan, 2003; Rhodes & Kelley, 2005; Koriat & Goldsmith, 1996; Koriat et al., 2008; Thomas & McDaniel, 2012). For example, Koriat and Goldsmith (1996) had participants answer general knowledge questions (e.g., What was the name of the first emperor of Rome?). The first step was a forced-report test: participants were first required to provide an answer to the question. Participants then rated their confidence in the accuracy of their provided answer. Finally, participants chose whether to report or withhold the answer they provided in the first step and participants earned money based on the accuracy of their responses on this free-report test. Reporting a correct answer would earn money, reporting an incorrect answer would lose the equivalent amount of money, and withholding an answer would have no effect on earnings.

Post-retrieval monitoring was highly accurate: confidence judgments were strongly correlated with answer accuracy on the forced-report test. Confidence judgments were also associated with post-retrieval control decisions: Average confidence ratings were significantly higher for items that participants chose to report than for items that participants chose to withhold on the free-report test. Critically, Koriat and Goldsmith's (1996) results suggest that post-retrieval monitoring and control are means by which memory accuracy can be improved. Forced-report test accuracy was calculated as the proportion of all items for which a participant generated a correct answer on the forced-report test. For instance, if a participant answered 50 questions and was correct on 25 of these questions, forced reported accuracy would be 50% (i.e., 25/50). Free-report test accuracy was calculated as the proportion of reported answers that were correct. For instance, if that same participant volunteered an answer on 30 questions and was correct on 25, free report accuracy would be 83% (i.e., 25/30). Free-report test accuracy was significantly higher than forced-report accuracy, suggesting that participants effectively monitored the accuracy of their retrieval output and controlled their memory reports accordingly.

In sum, previous research suggests that there are two stages to retrieval: front-end retrieval processes control the information that is produced during a memory search whereas back-end retrieval processes determine which information is selected and reported on the test. Consistent with this broader model of retrieval, SEED suggests that making errors while learning can enhance memory for the correct answer through both improved front-end and back-end retrieval processes. The precise mechanisms by which SEED improves front-end and back-end retrieval processes is described next.

### ***Strategic Encoding: Improving Front-End Retrieval Processes***

Previous research suggests that one route to accurate memory performance is through front-end retrieval processes, by narrowing the search set to only include relevant information (e.g., Jacoby et al., 2005). The strategic encoding component of SEED pertains to how making errors improves subsequent memory for the correct answer by affecting front-end retrieval processes. Broadly speaking, the strategic encoding component of SEED proposes that making an error enables individuals to better encode the correct answer once it is presented compared to errorless learning. Encoding the correct answer well initially will improve subsequent test performance by narrowing the search set, increasing the chances that only the correct answer will be produced as a possible answer, and not related information nor the original error.

**The Impetus Behind Strategic Encoding.** According to associative models of memory such as the Search of Associative Memory model (SAM; Raaijmakers & Shiffrin, 1981; Raaijmakers, 2003), when a cue is associated with multiple pieces of information, the probability that a target piece of information will be retrieved depends on two factors: 1) the size of the search set, i.e., the number of pieces of information associated with a cue and 2) the strength of the cue-target association relative to the association strengths between the cue and the other information in the search set. A target is more likely to be retrieved when a cue is associated with fewer additional targets rather than more additional targets (Figure 4a vs. 4b; e.g., Tulving & Pearlstone, 1966). A target is also more likely to be retrieved when it is strongly associated with the cue relative to the other associates of the cue (Figure 4c vs. 4b; e.g., Raaijmakers & Shiffrin, 1981; Raaijmakers, 2003). Therefore, learning conditions that selectively strengthen the cue-target association will improve subsequent memory performance by increasing the chances the target is recalled. SEED proposes that making an error is an example of a learning condition that selectively strengthens the cue-target association upon studying.

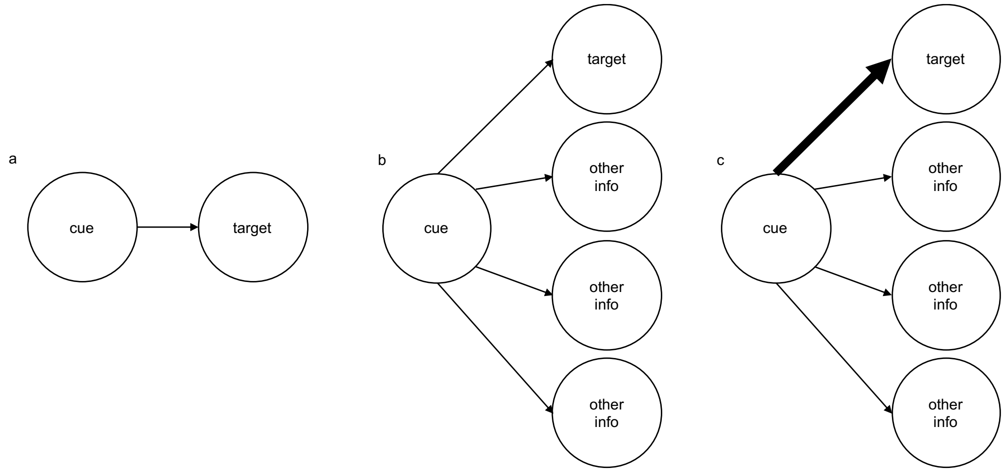


*Figure 4*. Schematics of different scenarios of how information could be stored in memory, according to associative models of memory. The thickness of the arrows represents the strength of the association in memory. A cue could only be associated with the target (a) or multiple

pieces of information (b and c). The target is more likely to be recalled when the cue is associated with fewer (a) rather than more (b) pieces of information. The association strength can also vary between a cue and different pieces of information. The target is more likely to be recalled when it is more strongly associated with the cue relative to other pieces of information associated with the cue (c).

Associative models of memory can be applied to describe how errorful and errorless learning conditions may affect memory. Before an error is eventually produced, there is some pre-existing association between the question and the error—that is why the error is the response that initially comes to mind (Figure 5a). If a failed retrieval attempt is made and an error is produced, the association between the question and the error is strengthened by the nature of producing the error (Figure 5b; e.g., Rowland, 2014; Slamecka & Graf, 1978). This pre-existing question-error association remains unchanged, though, in the errorless learning condition because the error is not produced (Figure 5e). Once the correct answer is encoded, the question becomes associated with multiple pieces of information, namely, the error and the answer (Figure 5c).

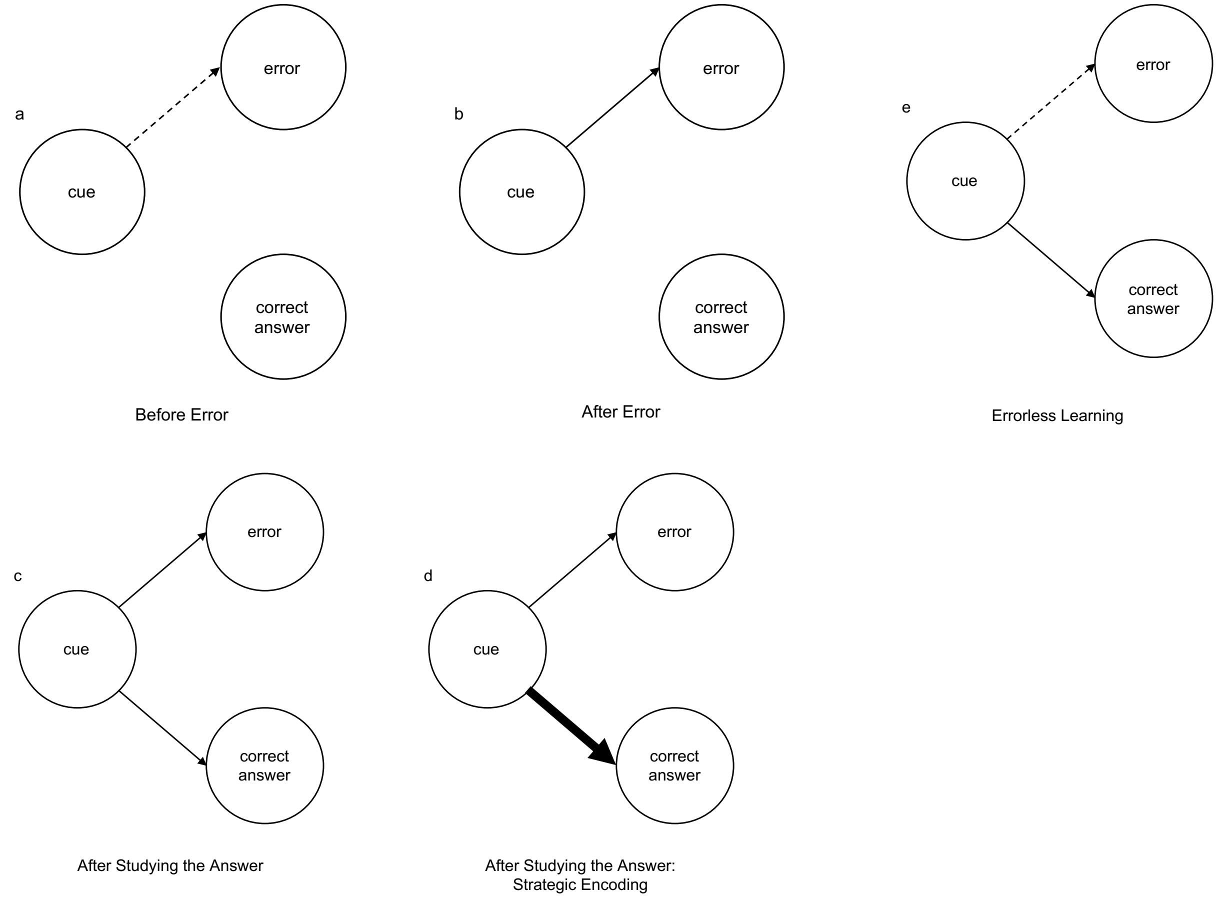


*Figure 5.* The strength of the association among a question (Q), error (E), and the correct asnwer (A). The thickness of the line indicates the strength of the association, with dashed lines representing the weakest association. Depicted are the associations before an error is produced (a), after an error is produced (b), after the answer is studied (c), after the answer is studied in a more strategic fashion (d), and under errorless learning conditions (e).

According to associative models of memory, producing an error should therefore interfere with, not facilitate, memory for the correct answer because the question-error association is relatively stronger under errorful rather than errorless learning conditions (Figure 5c vs. 5e). How can producing an error enhance learning? SEED invokes an idea that I refer to as *strategic encoding*.

**Errorful Learning Enables Strategic Encoding.** I define strategic encoding on a continuum as the degree to which an individual can effectively adapt their studying of the correct answer to meet their personalized learning needs for that item. SEED proposes that generating an error engenders more strategic studying of the correct answer once it is presented relative to errorless learning.

There are at least three related types of strategic encoding: attention, effort, and information. First, relative to an errorless condition, generating an error may lead to increased attention to the answer during feedback (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009). Consistent with the attention account, Butterfield and Metcalfe (2006) examined participants' abilities to detect tones that were presented during feedback after errors. Participants were less likely to detect the tones during feedback after high-confidence errors than low-confidence errors, suggesting that individuals can adjust their attention as needed on an item-by-item basis. Similarly, making an error may enable learners to better allocate their study time by spending more time studying the more difficult items (Soderstrom & Bjork, 2014; Yang, Potts, & Shanks, 2017). Indeed, individuals tend to devote more time to studying an answer after making a mistake than after correctly producing the answer (Soderstrom & Bjork, 2014).

Attention and study time can be considered quantitative measures of encoding that capture how much a participant is studying the answer. Strategic encoding may also involve qualitative shifts in the nature of the information that is encoded (e.g., Tversky, 1973). Previous research suggests that individuals can effectively shift the qualitative nature of how they encode information in order to maximize learning. For example, Finley and Benjamin (2012) had participants learn six lists of different word pairs; each list contained related and unrelated word pairs. For each list, participants studied each pair in the list at their own pace before taking a test on the target words in the list. For the six lists, participants completed three cued recall tests (e.g., cue-???) and three free recall tests, which involved recalling the targets unprompted. Across the six lists, the test format alternated between cued-recall and free recall so that each participant could experience both cued and free recall tests. Participants were informed before studying each list which type of test they would be taking so that they could adjust their study strategies accordingly.

Cued recall tests assess one's memory for the cue-target association, which are inherently weaker for unrelated pairs than related pairs. Therefore, unrelated word pairs are more difficult to learn for a cued recall test than related word pairs and require more study time. Indeed, participants spent more time studying unrelated word pairs than related word pairs across all three cued-recall lists. Participants began using a similar strategy for learning word pairs for the free recall tests. However, word pair relatedness had a much smaller effect on free recall test performance than cued recall test performance. Throughout the experiment, as participants gained experience with free recall tests, participants changed strategies; they stopped differentiating as much and study times became more similar for related and unrelated word pairs across the three free recall lists. The study time data provides strong evidence that individuals are able to strategically adjust how they study in order to maximize learning on a particular task.

Finley and Benjamin (2012) found that participants' self-reported study strategies corroborated patterns in study time data. At the end of the experiment, participants were asked to describe how they studied the items in preparation for cued and free recall tests. Participants' self-reported study strategies were coded as being more relational (i.e., focusing on the cue-target association through, say, interactive imagery) or more target focused (e.g., imaging just the target word). Participants reported adapting the qualitative nature of their encoding strategies to meet task demands; for instance, participants reported using more relational strategies for cued-recall tests and more target-focused strategies for free recall tests. In short, in order to maximize learning, individuals can both strategically adapt how much they study and the qualitative nature of how they study (e.g., the type of information they focus on).

There is reason to suspect that generating an error, in particular, can encourage qualitative shifts in the nature of the information that is encoded. Pyc and Rawson (2010) had participants learn Swahili translations (e.g., wingu-cloud) and asked them to report a mediator to help decode the Swahili word (e.g., wingu→wing→cloud). After an initial study phase, participants either restudied the pair and reported a mediator again or made a retrieval attempt, received feedback, and reported a mediator again. Participants were instructed that they could provide any mediator and did not have to report their original mediator. After approximately 95% of restudy trials, participants reported the same mediator as the one they had generated during the initial study phase. That is, mediator changes only occurred on 5% of trials. Mediator changes occurred after about 10% of successful retrieval trials. Critically, after participants generated an error during retrieval practice, they shifted to using a new mediator on approximately 30% of trials. Thus, errors led to qualitative changes in how the information was encoded (see also, DeWinstanley & Bjork, 2004).

SEED proposes that making an error subsequently engenders more strategic encoding of the correct answer than only studying the correct answer. One way in which making an error could engender more strategic encoding is by providing the learner with more diagnostic information about how to encode the correct answer. For instance, item difficulty may signal a need for changes in encoding. Attempting to retrieve the answer and making an error reveals how difficult the item is for the participant to learn. Therefore, making mistakes enables participants to spend the most time on the most difficult to learn items (Koriat, 1997, 2008; Lovelace, 1984; Soderstrom & Bjork, 2014). Making an error could also provide more diagnostic information about how effective one's current study strategies are than errorless learning (e.g., DeWinstanley & Bjork, 2004; Finley & Benjamin, 2012; Pyc & Rawson, 2010). For example, as described previously, Pyc and Rawson (2010) found that participants shifted their mediators after an error but not after restudying the pair.  Thus, making a mistake revealed to participants that their qualitative way of encoding the information was not optimal.

In sum, relative to errorless learning, making an error can provide diagnostic information about the item, which enables individuals to more effectively adapt how and how much they study the correct answer. Experiment 1 tested the hypothesis that making an error encourages shifts in encoding strategies.

Regardless of the particular mechanism—whether it be through increased attention and effort or a shift in the nature of the information that is encoded—SEED posits that making an error can improve encoding of the correct answer, thereby strengthening the question-answer association relative to the question-error association (Figure 5d vs. 5c). Strengthening the error-answer association during encoding will have down-stream consequences on future front-end retrieval processes. With a strengthened question-answer association, it is more likely that only the correct answer and not the error or other related ideas will be generated on a subsequent test (e.g., Raaijmakers & Shiffrin, 1981). Thus, SEED differs from the errors-as-mediators account in terms of front-end retrieval processes. The errors-as-mediators account suggests that errors enhance memory for the correct answer because both the original error and answer come to mind on a subsequent memory test. In contrast, SEED suggests that errors enhance memory for the correct answer because the correct answer is retrieved to the exclusion of the original error.

**Accounting for the Effects of Materials on Learning from Errors.** SEED is consistent with a much broader literature on encoding and retrieval. In order for SEED to be a viable model of error correction, though, SEED must be able to account for typical findings in the literature regarding learning from errors. Most of the evidence for the errors-as-mediators hypothesis comes from the finding that errorful learning tends to enhance memory for correct answers from semantically rich materials (e.g., related word pairs, trivia questions) but not for more semantically impoverished materials (e.g., unrelated word pairs, word stem completions; Bridger & Mecklinger, 2014; Cyr & Anderson, 2015; Grimaldi & Karpicke, 2012; Huesler & Metcalfe, 2012; Knight et al., 2012; Kornell, 2014).

SEED accounts for this finding by proposing that the degree to which generating an error enhances subsequent encoding depends on the degree to which the materials afford more strategic encoding—i.e., the degree to which more time and attention will enhance learning or the degree to which a learner can shift their study strategies to a qualitatively more effective one. Trivia questions afford far more strategic encoding opportunities (e.g., relating the answer to prior knowledge, explaining or justifying the answer) than, say, word stems (e.g., br_______ - brown). For semantically impoverished materials, it may be difficult for learners to identify more effective encoding strategies than maintenance rehearsal and additional time spent studying may yield little benefit to learning (Nelson & Leonesio, 1988).

Thus, in contrast to the errors-as-mediators hypothesis, SEED suggests that it is the degree to which the materials afford strategic processing of the correct answer—and not item relatedness *per se*—that causes learning from errors. Therefore, SEED predicts that errorful learning would improve memory for less semantically rich materials such as unrelated word pairs (e.g., door-shoe) in comparison to errorless learning if participants employed more effective encoding strategies of the correct answer, such as interactive imagery (e.g., Bower, 1970).

**Summary of Strategic Encoding.** In sum, an outline of the strategic encoding component of SEED is as follows:

a. Compared to errorless learning, making an error leads to heightened attention to the answer, increased effort to encode the answer, and/or a shift in the qualitative nature of what information is encoded and how.
b. The degree to which errorful learning conditions enhance memory for the answer on a later test depend on the degree to which the materials and task conditions afford effective encoding strategies.
c. Compared to errorless learning, strategic encoding of the answer during feedback improves memory for the answer by narrowing the scope of the memory search on the test to include the answer and not the error or other related ideas.

Experiments 1 and 2 will test the different predictions of the strategic encoding component of SEED.

### *Episodic Discrimination: Improving Post-Retrieval Monitoring*

The strategic encoding component of SEED is based on the finding that one route to accurate memory is to narrow the search set to include the correct answer, but not the original error or other related information (e.g., Jacoby et al., 2005). When strategic encoding of the correct answer is sufficiently high—as is predicted to be the case with semantically rich materials such as related word pairs—then front-end retrieval processes should produce the correct answer and not one's original error. If strategic encoding of the correct answer is not sufficiently high—perhaps because the materials are not semantically meaningful enough as with unrelated word pairs—then the error and answer will be associated with the question to a similar degree (Figure 5c). Therefore, the error and answer would have a similar likelihood of being included in the search set (Raaijmakers & Shiffrin, 1981). However, back-end retrieval processes (i.e., post-retrieval monitoring) can beget accurate recall even when multiple candidate answers are produced (e.g., the correct answer, the error, and other pieces of information). Therefore, the episodic discrimination component of SEED is based on the finding that another route to accurate memory performance is through accurate post-retrieval monitoring (e.g., Koriat & Goldsmith, 1996).

According to SEED, episodic memory is used to differentiate the correct answer from other candidate answers that are produced during the front-end memory search, including one's original error. To resolve the conflict among multiple candidate answers, episodic memory can be used to think back to the event in which the error was made and the correct answer was presented. If the study episode in which the error was made and correct answer was presented can be mentally reinstated, then learners will be able to consciously recall the details associated with the episode, including the source of the information encountered (e.g., self-generated error vs. experiment-presented correct answer; e.g., Johnson, Hashtroudi, & Lindsay, 1993). That is, another route to accurate final test performance is to rely on episodic memory to reinstate the learning context and memory for the source (i.e., source memory) to determine which piece of information was the correct answer. If episodic retrieval fails, though, it is more likely that the initial error is mistakenly reported as the correct answer.

**The Impetus Behind Episodic Discrimination.** Evidence for the role of episodic memory in resolving competition among multiple responses comes from the proactive interference literature. The typical proactive interference paradigm shares many similarities with the standard error correction paradigm. In the proactive interference paradigm, participants learn an association (A-B) and then later learn new conflicting (A-D) or non-conflicting (C-D) associations. Proactive interference refers to the finding that memory for the new information (D) is generally worse in the conflicting than the non-conflicting condition (e.g., Barnes & Underwood, 1959). Research suggests that a key predictor of whether old information (A-B) interferes with learning new information (A-D) depends on the degree to which the learning episodes (A-B and A-D) can be mentally reinstated (Jacoby & Wahlheim, 2013; Jacoby, Wahlheim, & Kelley, 2015; Jacoby, Wahlheim, & Yonelinas, 2013; Putnam, Wahlheim, & Jacoby, 2014; Wahlheim & Jacoby, 2013; Wahlheim, 2014, 2015).

For example, Wahlheim (2015) had participants study a list of word pairs (A-B; e.g., pearl-harbor) and then take a test on those pairs (e.g., pearl-???). Next, participants learned a new list of word pairs that contained items that repeated from list one (A-B, A-B pairs, e.g., pearl-harbor), changed from list one to list two (A-B, A-D pairs, e.g., pearl-jewelry), or were control items that had no relation to any items on list one (C-D pairs, e.g., baby-cute). The criterial test of interest was a cued recall test for list two pairs. In addition to completing a cued-recall test for the list two pairs, participants were also asked for on the final test whether an item repeated from list one to list two, changed from list one to list two, or was new to list two. This question probed participants' ability to use episodic memory to reinstate the study events of list one and list two. A mixture of opposing memory effects was observed on the final test. Relative to the control items (i.e., the C-D pairs), memory was worse for the changed pairs (i.e., the A-B, A-D pairs) when participants did not recall that the pair had changed from list one to list two. In contrast, memory was better for the changed pairs than the control pairs when participants recalled that the pair had changed. Thus, conflicting information need not produce proactive interference as long as the study episodes can be mentally reinstated and differentiated. In an analogous fashion, generating an error (cue-error) need not interfere with remembering the correct answer that was presented after making the error (cue-answer) as long as the study episode in which the error and answer were encountered can be recalled.

**Errors Enhance Episodic Memory.** A key assumption of SEED is that, in comparison to errorless learning, making an error will create a stronger episodic memory trace for the context surrounding when the error was made and the correct answer that was presented during learning. This episodic memory trace will include the question and error and can include the correct answer if feedback is immediately provided. Therefore, the test question can serve as an effective cue to mentally reinstate the study episode and the learner can use source-memory to differentiate the correct answer from the other information that comes to mind (e.g., Johnson et al., 1993).

In contrast, SEED predicts that merely studying correct answers will produce weaker episodic memory for the study experience. On the test, the question will be presented, but will be a less effective cue for reinstating the original study episode. If multiple candidate answers have come to mind on the test but the original study episode cannot be mentally reinstated, the learner will not be able to use source memory to determine which of the candidate answers had been previously presented as the correct answer (e.g., Johnson et al., 1993).

The assumption that errorful learning enhances episodic memory for the study context is based on prior research suggesting that retrieving previously studied information (Akan, Stanley, & Benjamin, 2018; Bishara & Jacoby, 2008; Chan & McDermott, 2007; Cyr & Anderson, 2012; Karpicke, Lehman, & Aue, 2014; Rowland & DeLosh, 2014; Wahlheim, 2015) and making mistakes (Anderson & Craik, 2006; Fazio & Marsh, 2009) enhances episodic memory. For example, Fazio and Marsh (2009) found that participants remembered not only the correct answers but also the color in which the answers had been presented better after making a surprising mistake compared to a less surprising mistake. Thus, surprising mistakes enhanced memory for not only the content participants were tasked to learn, but also the episodic details of the task.

**Summary of Episodic Discrimination.** In sum, an outline of the episodic discrimination component of SEED is as follows:

a. On a test, multiple candidate answers may be produced during the memory search. The candidate answers could include the correct answer, related ideas and concepts, and an error if one was produced.
b. Post-retrieval monitoring must be used to select the correct answer from the other candidate answers. Episodic memory—or mentally reinstating the initial learning phase—is one way in which the correct answer can be identified.
c. Producing an error improves episodic memory for the event in which the mistake was made and the correct answer was subsequently presented. Therefore, post-retrieval monitoring will be better under errorful than errorless learning conditions.

In its current form, SEED proposes episodic memory will effectively differentiate the correct answer from other candidate answers only if multiple candidate responses are activated during the initial memory search (although the two stages need not be independent; Kelley & Sahakyan, 2003). Therefore, SEED predicts that post-retrieval monitoring will be more critical for accurate memory performance under conditions that make strategic encoding less likely (e.g., learning unrelated word pairs). As described in the next section, one goal of Experiment 2 is to elucidate such post-retrieval monitoring processes.

## Present Experiments

SEED is based on a two-stage model of retrieval involving front-end control over what is retrieved from memory and post-retrieval monitoring of which piece of retrieved information is the correct answer. The errors-as-mediators account of error correction suggests that errors are used to help retrieve the correct answer on subsequent tests via semantic memory. In contrast, SEED suggests that errorful learning improves memory by narrowing the scope of subsequent memory searches (to exclude the error) or facilitating episodic retrieval of the original study episode.

Experiment 1 tested the hypothesis that making an error engenders more strategic encoding of the correct answer than errorless learning (at least with related word pairs). Experiment 2 will examine both the strategic encoding and episodic discrimination components of SEED by investigating how errorful and errorless learning affect the front-end and back-end retrieval processes involved with recalling the correct answer on a later test. In doing so, Experiment 2 will also provide evidence that can differentiate SEED and errors-as-mediators accounts of error correction.

# CHAPTER 2 – Experiment 1

Experiment 1 focused on the strategic encoding component of SEED. SEED predicts that when an error is made the correct answer can subsequently be encoded more strategically, thus enhancing long-term learning. Previous research has examined encoding strategies by asking participants to self-report their encoding strategies (e.g., repetition, imagery, generating sentences, etc.) either concurrently while studying or retrospectively once the learning task has been completed (e.g., Dunlosky & Hertzog, 1998, 2001; Finley & Benjamin, 2012; Pyc & Rawson, 2010). One drawback of using concurrent self-report measures of encoding strategies is that asking participants about their encoding strategies may cause participants to change how they study and affect learning as a result (Dunlosky & Hertzog, 2001; Fox, Ericsson, & Best, 2011; for guidance on effective use of think-aloud protocols, see Ericsson & Simon, 1980, 1998). For example, Pyc and Rawson (2010) had participants repeatedly study Swahili-English translations (e.g., wingu-cloud) and asked participants report the mediator that they used to encode each translation as they studied it (e.g., wingu→wing→cloud). Pyc and Rawson tracked changes in the mediators for a given translation as evidence of strategic shifts in encoding. However, participants may not have used mediators to encode the Swahili-English translations if they had not been asked to report mediators.

## Study Times as a Measure of Strategic Encoding

Experiment 1 collected objective behavioral measures that would be indicative of strategic encoding. Specifically, study times have been shown to reflect study strategies (for a review, see Kornell & Finn, 2016). For example, individuals tend to spend the most time studying the most difficult or least well learned information (e.g., Koriat & Bjork, 2006;

Soderstrom & Bjork, 2014; Son & Metcalfe, 2000; Yang, Potts, & Shanks, 2017). However, individuals can change how they allocate study time to respond to time pressure (e.g., Metcalfe & Kornell, 2003, 2005), incentives (e.g., Ariel & Dunlosky, 2013; Ariel, Dunlosky, & Bailey, 2009; DeLozier & Rhodes, 2015), and task experience (DeWinstanley & Bjork, 2004; Finley & Benjamin, 2012). Thus, study time can be a useful indicator of how study strategies vary under different learning conditions.

One indicator of strategic encoding is that study times for a given item are idiosyncratic to each participant. That is, strategic encoding frequently produces variability in how long different participants study a given item.  Such variability in study times reflects privileged access to the personal aspects of the items, including recent exposure to one of the words in the pair, personal experiences, and prior knowledge (Koriat, 1997, 2008; Lovelace, 1984). For example, a participant who recently took his toddler to swim lessons may spend less time studying the pair *swim-float* than other participants because the image of his child floating at the swim lesson may come to mind quickly and easily. Idiosyncrasies in individuals' study times are not merely random variations in study behavior. Previous research has demonstrated that idiosyncrasies in study times are indeed strategic because they lead to better learning. Allowing individuals to control how long they study each item produces better learning than when an experimenter determines the study time for each item, even when study times are equated (de Jonge, Tabbers, Pecher, Jang, & Zeelenberg, 2015; Mazzoni & Cornoldi, 1993; Nelson, Dunlosky, Graf, & Narens, 1994; Tullis & Benjamin, 2011; but see, Koriat, Ma'ayan, & Nussinson, 2006).

For example, Tullis and Benjamin (2011) had participant study a list of 80 words. Half of the participants were given control over how long to study each word. Therefore, the study time for each word and the total amount of time spent studying all 80 words was allowed to vary across participants in this self-regulated condition. For the other half of the participants, the computer controlled the study duration for each item. Each participant in the computer-controlled condition was yoked with a participant in the self-regulated condition. The participant in the computer-controlled condition was given the same total amount of time to study all 80 words as their counterpart in the self-regulated condition. However, for the participant in the computer-controlled condition, the amount of time to study each individual word was determined by the normative difficulty of the word; the computer allotted more time for words that previous research had shown were more difficult to remember. On a subsequent recognition memory test, participants in the self-regulated condition significantly outperformed participants in the computer-controlled condition.

These results suggest that the normative difficulty of an item is not an optimal indicator of how a piece of information should be studied. Instead, the amount of time allocated to studying a word should be adjusted for each participant to how difficult that particular word is for the individual participant. The fact that participants in the self-regulated condition learned more than participants in the computer-controlled condition suggests that item difficulty varies across participants and, critically, participants have accurate insight into what this difficulty is. That is, participants can determine which items are more difficult for them (independent of how difficult the item is, on average) and effectively adjust their study times to be more idiosyncratic accordingly.

## Predictions

A key assumption of SEED is that, relative to errorless learning, producing an error improves access to and the diagnostic quality of information about idiosyncratic item difficulty.

Therefore, making an error allows individuals to effectively adapt how they subsequently encode the correct answer. The more participants strategically adapt their encoding of a given item, the more study time for a given item should be unique to the individual participant. Therefore, Experiment 1 tested the prediction from SEED that study times for correct answers should be more idiosyncratic in the errorful than the errorless condition. The Results section describes how the idiosyncrasy of study times was computed based on methods used in previous research (Koriat, 1997, 2008; Lovelace, 1984).

## Methods

### *Participants*

The target sample size was 130 participants, which was selected to have 80% power to with alpha at .05 to detect an effect size of 0.5 for an independent samples *t*-test. The effect size of 0.5 is a conservative estimate based on previous research using highly similar materials and procedures (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Kornell, 2014; Kornell, Hays, & Bjork, 2009). More than 130 participants were collected to account for attrition or non-compliance with the task instructions, yielding a final sample size of 137.

One-hundred and thirty-seven participants were recruited using Prolific and were paid $5.00 for completing Experiment 1. The experiment took an average of 22.79 and 33.39 minutes to complete in the Errorless and Errorful conditions, respectively. One participant in the Errorful condition was excluded for not following the task instructions by failing to enter guesses on at least 75% of the guess trials. Among the remaining 136 participants, 68 were randomly assigned to the Errorful condition (mean age = 28.29 years, 21 female, 45 male, 1 non-binary, mean education = 14.69 years, all fluent English speakers) and 68 were randomly assigned to the

Errorless condition (mean age = 30.62 years, 23 female, 44 male, 1 preferred not to say, mean education = 15.03 years, all fluent English speakers).

### *Materials*

The materials were 60 related word pairs (e.g., swim-float; see Appendix A for full list of materials) selected from the norms of Nelson, McEvoy, & Schreiber (1998), with a weak forward association between 0.05 and 0.054 (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Kornell, 2014; Kornell et al., 2009). For example, when presented with the word *swim*, only approximately 5% of individuals report *float* as the first word that comes to mind. Given these norms, when presented with the first word in the pair, approximately 5% of participants were expected to generate the second word in the pair. Therefore, nearly all of participants' guesses were expected to be errors.

### *Design and Procedure*

Participants were randomly assigned to an errorful or errorless learning condition (Figure 6) and completed an initial learning phase, a distractor task, and a final test.

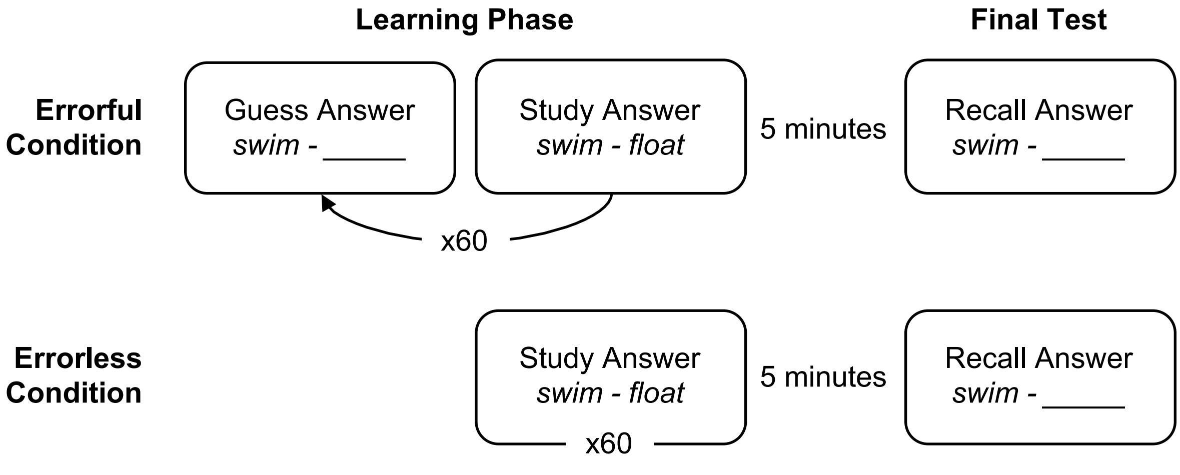


*Figure 6.* Experiment 1 design. Participants were randomly assigned to either the errorful or errorless learning condition. Participants studied 60 weakly related word pairs, solved arithmetic problems as a 5-minute distractor task, and then took a cued recall test. However, in the errorful

learning condition, participants guessed the answer before studying the correct answer for each pair in the learning phase.

During the learning phase in the errorful learning condition, participants were shown the cue (e.g., *swim*) and had as long as they needed to guess the answer (e.g., a participant could guess *pool*). Next, the correct answer (e.g., *swim-float*) was presented and participants studied the correct answer for as long as needed before proceeding to the next pair. In the errorless learning condition, participants only studied the correct answers (e.g., *swim-float*), which were presented one at a time, and participants studied the correct answer for as long as needed before proceeding to the next pair. After the learning phase, participants completed arithmetic problems for 5-minutes as a distractor task before the final test. On the final test, the cue was shown (e.g., swim) and participants attempted to recall the correct answer (e.g., float). Participants had as long as necessary to recall each answer and no feedback was provided on the final test.

Two random item orders were created and were counterbalanced across participants. That is, 34 participants in the errorful condition learned the items in order A and the remaining 34 participants in the errorful condition learned the items in order B. Similarly, 34 participants in the errorless condition learned the items in order A and the remaining 34 participants in the errorless condition learned the items in order B. The two item orders facilitated comparing study times during the learning phase across participants.

After the final test, participants completed a study strategy inventory (Table 1) in which they reported the strategies they used while studying the pairs (adapted from Karpicke et al., 2009).

Table 1

*Study Strategy Inventory*

| Which of the following strategies did you use to study the word pairs? Check all that apply. |
|---|
| Repeated the pair over and over again in your head |
| Made a mental image of the words in the pair |
| Created mnemonics such as rhymes or acronyms to connect the words in the pair |
| Came up with a sentence or a story to connect the words in the pair |
| Related the word pair to yourself |
| Acted out or imagined acting out the words in the pair |
| Tested yourself by covering up the second word in the pair |
| Other: |

## Results

### ***Guessing Accuracy***

Consistent with existing norms (Nelson et al., 1998), participants in the errorful condition guessed the correct answer on 5.5% of trials. Because the purpose of Experiment 1 was to examine how making errors affects study time, these trials were excluded from subsequent analyses.

### ***Final Test Performance***

A between groups *t*-test revealed that final test performance was significantly higher in the errorful ($M = 0.64$, $SD = 0.22$) than the errorless condition ($M = 0.49$, $SD = 0.25$), $t(134) = 3.72$, $p < .001$, $d = 0.64$.

### ***Study Time Allocation***

Across all items and participants, the mean study time for the correct pairs was 4.34 seconds, the median study time was 1.815 seconds, and the standard deviation of study times was 18.73 seconds. Individual trials with extreme study times were excluded from analyses of study time.[1] A study time for a given item was considered extreme if it was three standard deviations above or below the median study time for all items and all participants. Thirty-nine trials were excluded for being too long (i.e., longer than 60.52 seconds). No trials were excluded for being too short. Figure 7 depicts the distribution of study times in the errorful and errorless conditions, with the extreme trials excluded.

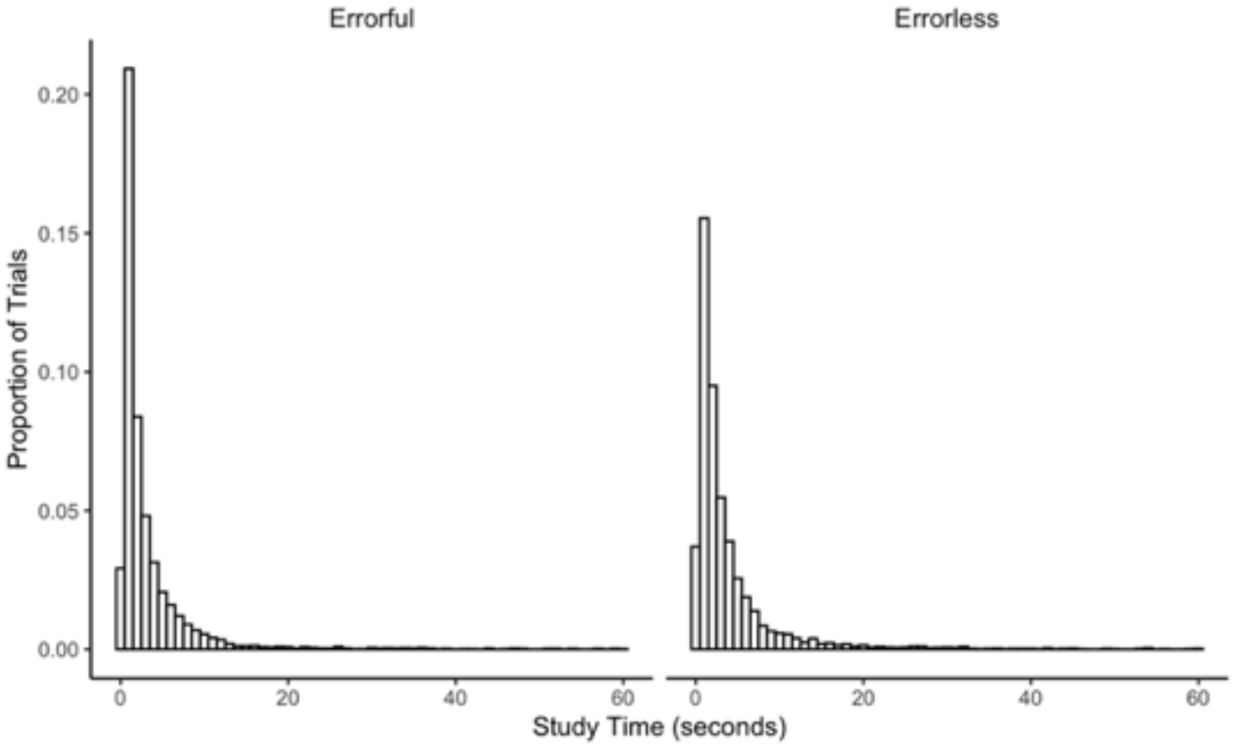


*Figure 7*. Relative frequency of study times in the errorful and errorless conditions. Each bar represents a range of one second.

No significant differences in mean study time emerged between the errorful (*M* = 3.47, *SD* = 3.97 seconds) and errorless conditions (*M* = 4.08, *SD* = 4.08 seconds), *t*(134) = 0.48, *p* = 0.63, *d* = 0.08. Similarly, no significant differences in median study time emerged between the errorful (*M* = 2.62, *SD* = 4.08 seconds) and errorless conditions (*M* = 3.00, *SD* = 3.43 seconds),

---

[1] The pattern of results was nearly identical when these trials were not excluded.

$t(134) = 0.59$, $p = 0.55$, $d = 0.10$. Thus, errorful learning increased memory for the correct answer, but did not increase the total time participants chose to study the correct answers.

The primary question of Experiment 1 was whether making an error engenders more strategic studying of the correct answer than errorless learning, as indicated by self-regulated study times during the learning phase. If participants were engaging in strategic study decisions, then their study times should be idiosyncratic. Therefore, to the degree that participants engage in strategic encoding, one participant's item-by-item final test performance should be better predicted by their own study times for each item than another participant's study times. To test this hypothesis, two study time-recall correlations were calculated: a personal and a yoked correlation.

For each participant, the personal correlation was calculated as the item-by-item ordinal Kruskal-Goodman gamma correlation (Nelson, 1984) between study time during the learning phase and recall on the final test. Gamma is a non-parametric measure of association, which is appropriate for examining the association between study times and recall on an item-by-item basis because no distributional assumptions must be made. Study times were not normally distributed (Figure 7) and recall was an ordinal variable (recalled or not recalled). A positive correlation would indicate that participants spent more time studying words that they correctly recalled on the final test whereas a negative correlation would indicate that participants spent less time studying words that they correctly recalled on the final test.

Each participant was then randomly matched with another participant in the same study condition (errorful or errorless) and the same item order counterbalance condition (to control for order effects in study time or final test performance). The yoked study time-recall correlations were calculated as the item-by-item gamma correlation between a participant's recall on the final test and their yoked participant's study times (Lovelace, 1984; Koriat, 1997, 2008). A 2 (study condition: errorful vs. errorless; between) x 2 (correlation type: personal vs. yoked; within) mixed-design ANOVA of study time-recall correlations was conducted. Strategic encoding would be reflected by a main effect of correlation type; specifically, participants should exhibit stronger personal than yoked correlations between study time and recall on the final test, reflecting insight into the idiosyncratic elements of the item. Therefore, because the SEED model posits that errorful learning engenders more strategic encoding of the correct answer, SEED predicted a significant interaction effect such that personal correlations would be stronger than yoked correlations, but more so in the errorful than errorless learning conditions.

Contrary to these predictions, there was little evidence of strategic encoding as measured by self-regulated study times. A 2 (study condition: errorful vs. errorless; between) x 2 (correlation type: personal vs. yoked; within) mixed-design ANOVA of study time-recall correlations revealed no main effect of correlation type, $F(1,134) = 0.86$, $p = .37$, = $\eta_p^2 = .006$, suggesting that one's own study times did not predict recall on the final test better than another randomly-selected participant's study times (Figure 8). Furthermore, although there was a significant interaction effect, $F(1,134) = 6.27$, $p = .02$, $\eta_p^2 = .05$, the effect was in the opposite direction predicted by SEED. Specifically, paired sample *t*-tests revealed that personal study time-recall correlations ($M = .11$, $SD = .26$) were significantly stronger than yoked correlations ($M =.02$, $SD = .18$) in the errorless condition, $t(67) = 2.36$, $p = .02$, $d = 0.23$. However, no difference emerged between personal ($M = -.04$, $SD = .20$) and yoked correlations ($M = .01$, $SD = .17$) in the errorful condition, $t(67) = 1.15$, $p = .26$, $d = 0.14$. Similarly, the personal study time-accuracy correlations were statistically larger in the errorless than the errorful condition, $t(134) = 3.71$, $p < .001$, $d = 0.63$.

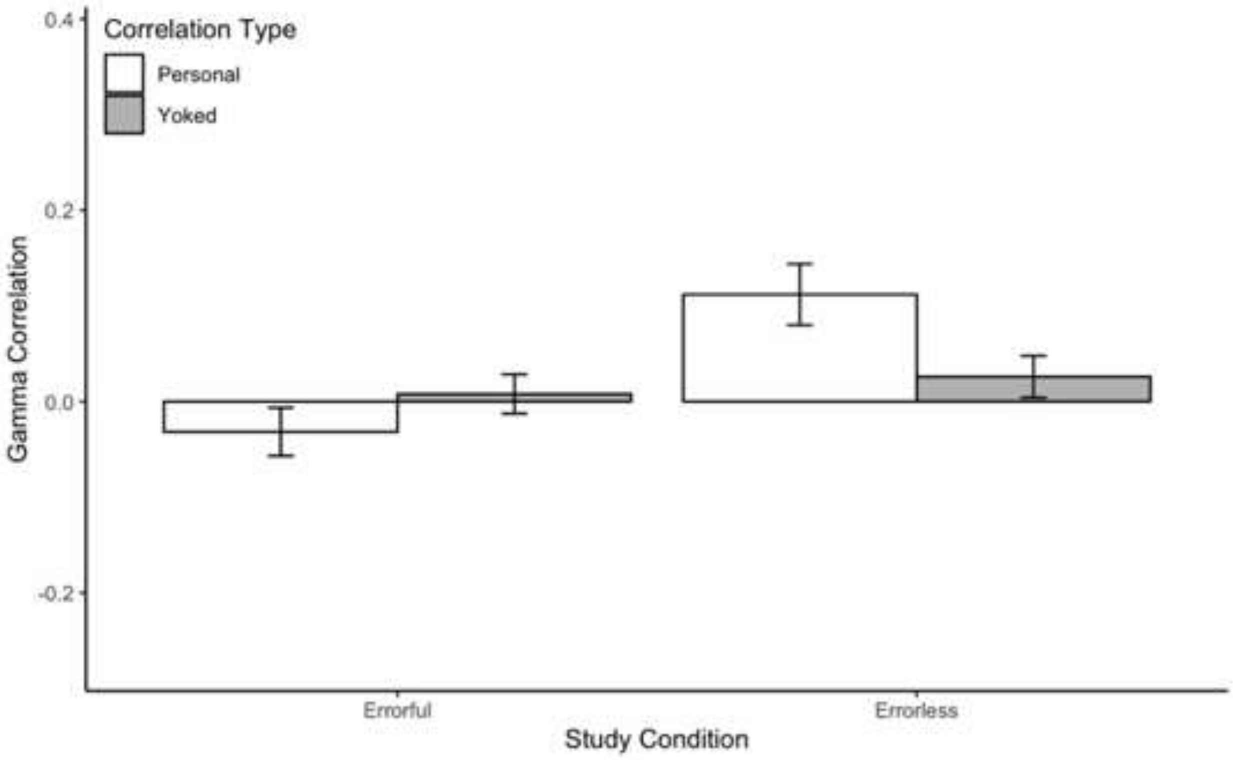


*Figure 8*. The personal and yoked item-by-item gamma correlations between study time and recall on the final test in the errorful and errorless conditions. Error bars represent ±1 standard error of the mean.

However, the statistically significant interaction and simple effects should be interpreted with caution. The magnitude of the observed correlations was small; there was little to no association between the time a participant spent studying a pair and whether the correct answer was recalled on the final test. In the errorful condition, the mean gamma correlation was not statistically different from zero, $t(67) = 1.47$, $p = 0.15$, $d = 0.18$. In the errorless condition, there was only a weak, albeit statistically significant, gamma correlation between study times and recall, $t(67) = 3.50$, $p < .001$, $d = 0.42$, suggesting that participants in the errorless condition were somewhat more likely to correctly recall the pairs they had spent more time studying. Thus, regardless of study condition, one's own study times predicted subsequent recall only weakly, if at all. Given the weak study time-accuracy correlations that were observed, any statistically significant differences in the correlations between the errorful and errorless conditions likely cannot account for the large differences in memory performance on the final test.

### ***Self-Reported Study Strategies***

After the final test, participants reported which study strategies they had used to encode the word pairs. The most common strategy that participants reported using was to repeat the pair over and over again. The least common strategies were to use mnemonics such as rhymes or acronyms to connect the words in the pair (Table 2, reported in order of frequency of use). However, study condition did not affect which strategies participants reported using. Exploratory analyses revealed no statistically significant differences between the errorful and errorless conditions in the proportion of participants reporting using any strategy.

Table 2

*Proportion of Participants Reporting Use of Study Strategies in Experiment 1*

| Study Strategy | Errorful Condition | Errorless Condition |
|---|---|---|
| Repeated the pair over and over again in your head | 0.44 | 0.50 |
| Made a mental image of the words in the pair | 0.45 | 0.48 |
| Related the word pair to yourself | 0.40 | 0.28 |
| Came up with a sentence or a story to connect the words in the pair | 0.13 | 0.20 |
| Acted out or imagined acting out the words in the pair | 0.05 | 0.22 |
| Tested yourself by covering up the second word in the pair | 0.07 | 0.16 |
| Created mnemonics such as rhymes or acronyms to connect the words in the pair | 0.12 | 0.07 |

In addition, 12 participants (8 in the errorful condition, 4 in the errorless condition) selected ‘Other’ on the study strategies survey and described their own strategies they had used. The strategies that they wrote in could be classified as one of the provided strategies on the survey (e.g., “recalled as I went” could be classified as testing one’s self or “tried to remember the first letter in my head” could be classified as using a mnemonic).

## Discussion

Consistent with prior research, Experiment 1 revealed that errorful learning significantly enhanced memory for correct answers on the final test relative to errorless learning. SEED posits that, in comparison to errorless learning, making an error provides the individual with more diagnostic information regarding how difficult the item is for them to learn (Koriat, 1997, 2008; Lovelace, 1984; Soderstrom & Bjork, 2014) or how effective one’s current way of thinking about the item is (Pcy & Rawson, 2010). According to SEED, generating an error while learning therefore enhances subsequent encoding of the correct answer by enabling individuals to strategically adapt their attention (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009; Peterson & Wissman, 2020), study time (Soderstrom & Bjork, 2014), and/or qualitative encoding strategies (e.g., Pyc & Rawson, 2010; deWinstanley & Bjork, 2004) to meet their study needs for each item.

Experiment 1 used study times as an indicator of strategic encoding processes (e.g., Koriat et al., 2006; Soderstrom & Bjork, 2014). Specifically, Experiment 1 tested the prediction of SEED that an individual’s item-by-item study times should be more predictive of which items are later recalled than another randomly yoked participant’s study times, and more so in the errorful than the errorless condition. Contrary to this prediction, Experiment 1 revealed that the difference between the personal and yoked study time-accuracy correlations was larger in the errorless than the errorful condition. Therefore, Experiment 1 revealed no evidence that generating errors enhanced learning of the correct answer by enabling participants to strategically allocate more study time to the items that were more difficult to learn and less time to the items that were easier for them to learn.

In fact, study time was generally not predictive of learning in Experiment 1 at all. The correlation between study time and recall was negligible in both the errorless and errorful conditions, in contrast to previous research which has generally found that individuals spend more time on more difficult items, which they are less likely to recall on the final test (e.g., Koriat, 2008; Koriat et al., 2006; for a review, see Son & Metcalfe, 2002). Although there is no apparent explanation for why the association between study time and recall was so weak, the key conclusion from Experiment 1 is clear: how participants allocated their study time cannot explain the large benefit of errorful learning. One alternative explanation for the benefit of errorful learning is that it is not study time that affects learning, but rather, the efficiency of the encoding processes during that study time. That is, errorful learning may have led to a stronger memory representation of the correct answer using the same amount of study time via increased attention (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009), less mind wandering (Peterson & Wissman, 2020), or a qualitative shift to more effective encoding strategies (DeWinstanley & Bjork, 2004; Pyc & Rawson, 2010). On the final test, the correct answer would therefore be more likely to come to mind during the front-end stage of retrieval.

However, participants' self-reported study strategies do not necessarily support this efficient-encoding account. There were no differences in self-reported study strategies between the errorful and errorless conditions. In both conditions, participants used both effective and ineffective study strategies, consistent with the finding that individuals tend not to be perfectly aware of which study strategies are effective (e.g., Hartwig & Dunlosky, 2012; Morehead, et al., 2016). For example, many participants reported using mental imagery, an effective encoding strategy (e.g., Bower, 1970). However, the most commonly reported study strategy was repetition, even though repetition is a relatively ineffective learning strategy (e.g., Craik & Watkins, 1973). Similarly, few participants reported testing themselves by covering up the second word in the pair, even though retrieval practice is a highly effective learning strategy (e.g., Rowland, 2014).

It is possible, though, that errorful learning promoted more efficient encoding of the correct answers than errorless learning in a way that participants could not articulate on a single self-report survey question at the end of the experiment. For example, if generating errors enabled participants to pay more attention during the study phase and engage in less off-task mind wandering compared to errorless learning (Peterson & Wissman, 2020), participants would not have been able to express this in the study strategy survey because study condition was manipulated between participants. Generating errors could have also supported subsequent encoding of the correct answer by activating related words in the semantic network (Carpenter 2009, 2011), which is not necessarily a conscious process. In short, the results of Experiment 1 suggest that if generating errors enhances memory for the correct answer by improving encoding, it is likely not the amount of encoding, but rather, the nature of the encoding processes. However, Experiment 1 cannot identify the exact nature of this more efficient encoding.

Another explanation for why Experiment 1 revealed no evidence for strategic encoding in study times is that errorful learning does not enhance memory for the correct information by improving initial encoding. Instead, errorful learning may improve memory for the correct answers primarily through the back-end retrieval processes involved in selecting the correct answer on the final test. That is, the correct answer may be equally likely to come to mind on the final test in the errorful and errorless conditions but errorful learning may better enable participants to select the correct answer among the other possible answers that come to mind. Experiment 2 examined these encoding-based and retrieval-based accounts of how errorful learning enhances memory for correct answers relative to errorless learning.

# CHAPTER 3 – EXPERIMENT 2

Experiment 1 tested the strategic encoding component of SEED and focused on the initial encoding phase. In contrast, Experiment 2 tested both the strategic encoding and episodic discrimination components of SEED and focused on the final test. The design of Experiment 2 thus provided evidence to weigh the relative value of the SEED and errors-as-mediators accounts of error correction.  In addition, Experiment 2 investigated the role of feedback timing under errorful and errorless learning conditions.

## Effects of Errorful Learning on Retrieval Processes

SEED and the errors-as-mediators hypotheses make different predictions regarding the mechanism by which making errors enhances subsequent learning of the correct answer. The two accounts can be differentiated by examining front-end and back-end retrieval processes separately.

### *Learning from Errors: Front-End Retrieval Processes*

Both SEED and the errors-as-mediators account make predictions about how making errors affects the information that comes to mind on a later test, i.e., front-end retrieval processes. According to SEED, the consequence of strategic encoding is that during a later retrieval attempt, the search set is narrowed and the correct answer is more likely to be recalled because fewer other candidate answers come to mind. In contrast, the errors-as-mediators account does not predict a narrowed search set. Instead, on a later test, the error and correct answer should both come to mind. Therefore, SEED and errors-as-mediators differ primarily in terms of the proposed front-end retrieval processes involved with recall of the correct answer.

Thus, Experiment 2 contrasted SEED with the errors-as-mediators account of error correction using a four-step final test format that was developed to examine front-end retrieval processes (e.g., narrow vs. wide search set) and back-end retrieval processes (e.g., post-retrieval monitoring to select the correct answer among multiple candidates; Koriat & Goldsmith, 1996; Goldsmith & Koriat, 2007; Halamish et al., 2012; Thomas & McDaniel, 2012). On the four-step test, a cue was presented and participants were asked to 1) report all the candidate answers that come to mind, 2) select which of those candidates is the correct answer, 3) rate the likelihood that each candidate is the correct answer, and 4) decide whether to report their answer selected in step 2 and earn points for a correct answer or risk losing points for an incorrect answer. SEED predicts that because of strategic encoding during the learning phase, fewer candidate answers would come to mind on step 1 of the final test in the errorful than errorless condition. When the correct answer was reported in step 1, it would typically be the only candidate reported. In contrast, errors-as-mediators predicted that when the correct answer was recalled in step 1, the participant's original guess would be reported, too.

### *Learning from Errors: Back-End Retrieval Processes*

If the error and then the correct answer are retrieved on the final test as the errors-as-mediators account posits, then participants must be able to differentiate the error from the correct answer. Therefore, the errors-as-mediators account also implies differences in back-end retrieval processes between the errorful and errorless conditions. This prediction has never been previously tested. Thus, examining front-end and back-end retrieval processes separately in Experiment 2 tested the strategic encoding component of SEED, tested a novel prediction of the errors-as-mediators hypothesis, and provided important evidence as to whether SEED or the errors-as-mediators hypothesis is a more viable account of how making errors enhances subsequent learning.

A priori predictions of SEED about how post-retrieval monitoring would differ between the errorless and errorful conditions were less clear. SEED posits that accurate post-retrieval monitoring is critical for accurate memory when multiple candidate responses come to mind on the test. Because Experiment 2 used related word pairs, SEED predicts that strategic encoding should be sufficiently effective to narrow the search set to the correct answer to the exclusion of the original error and other related words, making post-retrieval monitoring less essential for identifying the correct answer. Therefore, measures of post-retrieval monitoring accuracy could be similar in the errorless and errorful with immediate feedback conditions. However, the nature of the final test asked participants to explicitly engage in post-retrieval monitoring that was relevant even if participants only produced the correct answer on step 1. On step 3, participants rated the likelihood that answer was correct and then decided whether to report or withhold it on step 4. Asking participants to explicitly engage in post-retrieval monitoring could have encouraged the use of episodic memory to mentally reinstate the study episode. In this case, SEED would predict more accurate post-retrieval monitoring in the errorful with immediate feedback condition than errorless condition. Therefore, Experiment 2 provided additional evidence regarding SEED, specifically, the role of episodic discrimination in learning from errors under conditions in which strategic encoding is expected (i.e., with related word pairs).

## Effects of Feedback Timing on Learning from Errors

Not only has errors-as-mediators been proposed as a mechanism for the benefits of errorful learning, the errors-as-mediators hypothesis has been used to account for another key finding in the error-correction literature, namely, the effects of delaying feedback on learning from errors (Kornell, 2014; Vaughn & Rawson, 2012). Although making errors can enhance learning, the benefits have been shown to depend upon the timing of feedback (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Kornell, 2014; Vaughn & Rawson, 2012). For example, Grimaldi and Karpicke (2012) had participants learn related word pairs (e.g., tide-beach) under errorful or errorless conditions. Participants either only studied the correct pair (tide-beach) or generated a guess for the second word in the pair (tide-????) before studying the correct answer (tide-beach) either immediately or after an approximately 10-minute delay. In the delayed feedback condition, participants generated their guesses for all pairs before studying the correct answers for all of the pairs. On a subsequent final test, memory for the correct answers was better after errorful learning with immediate feedback compared to errorless learning. However, errorful learning did not enhance final test performance relative to errorless learning when feedback was delayed.

In order to be a viable account of how one learns from errors, the errors-as-mediators hypothesis and SEED must be able to explain why delaying feedback would impair learning from errors. Therefore, in addition to the errorless and errorful with immediate feedback conditions used in Experiment 1, Experiment 2 also included an errorful condition in which feedback was delayed by several minutes. In this delayed condition, participants generated a guess for every pair before studying the correct answer for every pair.

The errors-as-mediators hypothesis suggests that when a guess is made and then the correct answer is presented, the cue, guess, and target are all activated in close temporal proximity, allowing for an association between the guess and correct answer to be formed, creating a mediated path (Figure 9a). In contrast, when a guess is made, but the feedback is delayed, the activation of the guess fades before the correct answer is presented (McNamara, 2014). Therefore, the guess and the correct answer are not activated at the same time and an association is less likely to be formed between the guess and answer (Figure 9b).

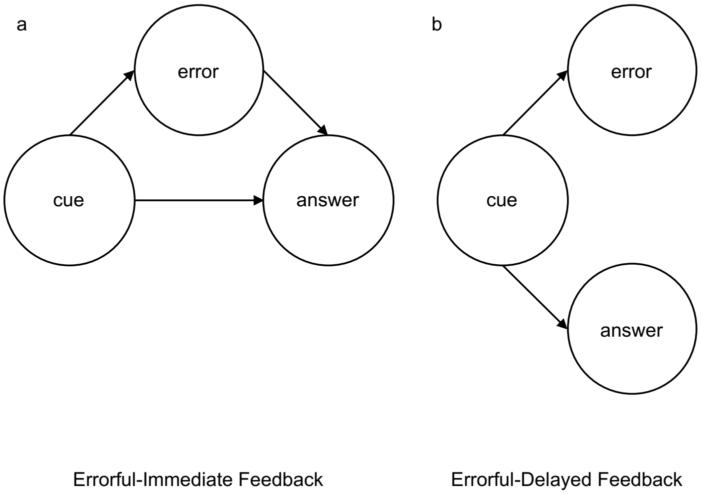


*Figure 9.* Possible associations in memory after errorful learning according to the errors-as-mediators hypothesis. When correct answer feedback is presented immediately after making a guess, a mediated pathway can be created from the cue to the guess to the correct answer and retrieval of the error can facilitate retrieval of the answer (a). When correct answer feedback is delayed, the guess and answer are not activated at the same time in memory and a mediated pathway is less likely to form and the cue-error association can interfere with retrieval of the correct answer (b).

On a subsequent test, the correct answer will be likely to be retrieved in the errorful condition when immediate feedback had been presented because the correct answer can be retrieved directly and through the cue→error→answer mediated retrieval route (Figure 9a). In contrast, the correct answer will be less likely to be retrieved if feedback had been delayed because when the cue activates the error, the error will not activate the correct answer, in turn. Instead, because the error and answer are not integrated into a single memory trace (Figure 9b), retrieving the error could interfere with retrieval of the correct answer (Vaughn & Rawson, 2012). In short, the mediator-based explanation posits that delaying feedback impairs learning because delayed feedback prevents the learner from forming a mediated retrieval pathway from the cue to the answer via one's original guess (Hays et al., 2013; Kornell, 2014; Vaughn & Rawson, 2012).

SEED offers an alternative explanation for why delaying feedback impairs learning from errors. According to SEED, making an error improves subsequent encoding of the correct answer through enhanced attention (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009), time (Soderstrom & Bjork, 2014), and/or a qualitative shift in encoding strategies (e.g., Pyc & Rawson, 2010). Thus, the strength of the cue-answer association becomes stronger than the cue-error association (Figure 10a).

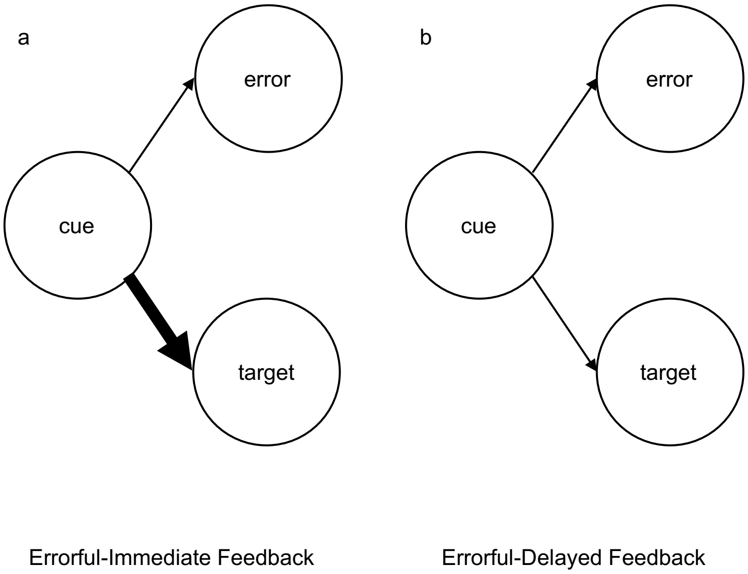


*Figure 10.* Possible associations in memory after errorful learning according to SEED. The thickness of the arrow indicates the strength of the association in memory. When correct answer feedback is presented immediately after making a guess, the target can be strategically encoded, creating a strong association between the cue and target (a). On a subsequent test, the target will therefore be more likely to be retrieved than the error. When correct answer feedback is delayed, the target cannot be as strategically encoded, leading to a much weaker association between the cue and the target (b). On a subsequent test, memory for the correct answer will be worse because the error and answer have a more similar probability of being retrieved; retrieving the error will cause interference for retrieving the answer.

However, delaying feedback may undermine the degree to which participants can appropriately modify their attention, time, or strategies on an item-by-item basis because of any activity that happens during the delay. Making an error can provide participants with diagnostic information about how difficult an item will be for them to learn and SEED posits that it is this diagnostic information that enables participants to adjust their encoding effort and strategies accordingly. In Experiment 2, participants made a guess for a pair and then multiple items intervened before the correct answer to the pair was presented. These intervening items could make this diagnostic information difficult to remember once the correct answer is presented several minutes later. Therefore, delaying feedback could make the experience of an error less informative for adapting subsequent encoding and therefore could undermine strategic encoding. Indeed, previous research suggests that the benefits of feedback are diminished when there is even a brief (e.g., 3 second) task separating when a participant answers a question and receives corrective feedback (Carpenter & Vul, 2011). Thus, SEED predicts that as a result of delayed feedback, the strength of the cue-answer association will be more similar to the cue-error association (Figure 10b) than if immediate feedback had been presented. On a subsequent test, the correct answer is more likely to be retrieved after errorful learning with immediate feedback than errorful learning with delayed feedback because of strategic encoding that selectively strengthens the cue-answer association in the immediate but not delayed feedback condition.

Given these alternative theoretical accounts, an additional purpose of Experiment 2 was to examine why delaying feedback impairs learning from errors. The four-step final test differentiated between front-end and back-end retrieval processes as a way to compare SEED and the errors-as-mediators accounts of the effects of feedback timing.

### ***Feedback Timing: Front-End Retrieval Processes***

Both errors-as-mediators and SEED make predictions regarding how feedback timing affects what information will come to mind on a final test, i.e., front-end retrieval processes. The errors-as-mediators hypothesis predicts that error→answer association is less likely to form if feedback is delayed than if feedback is immediate. Therefore, the errors-as-mediators hypothesis predicts that, in the immediate feedback condition, the correct answer should typically be retrieved when the original error is retrieved on step 1 of the final test. In contrast, in the delayed feedback condition, there should be a significantly lower probability that the correct answer is recalled when the original guess is recalled. That is, the errors-as-mediators hypothesis predicts that the error and answer should typically be recalled together in the immediate feedback condition, but only the error should be recalled in the delayed feedback condition.

SEED suggests that strategic encoding of the correct answer should selectively strengthen the association between the cue and the correct answer relative to the association between the cue and one's original guess. Therefore, on a subsequent memory test, it should be more likely that only the correct answer comes to mind in the immediate feedback condition compared to the delayed feedback condition. If delaying feedback interferes with strategic processing of the correct answer, then there should be more similar cue→guess and cue→answer association strengths, increasing the probability that both the original guess and correct answer will be retrieved. Therefore, in contrast to the errors-as-mediators hypothesis, SEED predicts that the error and answer should be more likely to both be retrieved as candidate answers on step 1 of the final test in the delayed condition compared to the immediate condition. In sum, the errors-as-mediators and SEED hypotheses make opposing predictions regarding how delaying feedback after making an error affects front-end retrieval processes. Experiment 2 tested these opposing predictions.

### *Feedback Timing: Back-End Retrieval Processes*

Both errors-as mediators and SEED predict that feedback timing should influence one's ability to select the correct answer among multiple candidates that come to mind on a test, i.e., back-end retrieval processes. According to the errors-as-mediators account, forming an association between the guess and the target may preserve information about the order in which words were encountered in the experiment (e.g., Jacoby et al., 2013; Wahlheim & Jacoby, 2013). Therefore, if the original guess and correct answer both come to mind on a later test, participants can use this order information to determine which word was more recently encountered and thus is the correct answer (Jacoby & Wahlheim, 2013). However, delaying feedback should interfere with forming an error→answer association because the error and answer are not active in close temporal proximity. Therefore, both the guess and correct answer may come to mind on the final test, but participants will not be as accurate in determining which word was encountered more recently and is therefore the correct answer. Thus, the errors-as-mediators account predicts impaired post-retrieval monitoring in the delayed feedback compared to the immediate feedback condition

In contrast, one interesting possibility based on SEED is that delaying feedback will actually improve post-retrieval monitoring accuracy. In both the immediate and delayed feedback, an error is made, rendering the episode in which the error was produced more memorable. However, in the delayed feedback condition of Experiment 2, participants made their guess for every pair before studying the correct answers for all of the pairs. Since temporal context changes as time passes (e.g., Howard & Kahana, 2002), the time at which the error was made and the time at which the correct answer was presented should be more distinct in episodic memory with delayed feedback than immediate feedback. Therefore, if one's original guess and the correct answer both come to mind on a test because delayed feedback impaired strategic encoding of the correct answer, the fact that the guess and answer were encountered in more distinct contexts may improve one's ability to select the correct answer as the word that was meant to be retrieved. Thus, examining back-end retrieval processes offers tests and comparisons of the errors-as-mediators and SEED accounts of why delaying feedback impairs learning.

In sum, both front-end and back-end measures of retrieval processes should reveal signatures of the errors-as-mediators account of learning from errors as well as the effects of feedback timing on learning from errors. Thus, the four-step final test format of Experiment 2 tested novel predictions of the errors-as-mediators account of error correction and test SEED. Furthermore, the errors-as-mediators and SEED accounts often make differing predictions regarding front-end and back-end retrieval processes. Therefore, Experiment 2 also contrasted the errors-as-mediators and SEED accounts as viable explanations of how making errors can enhance learning of correct information.

## Methods

### *Participants*

One hundred eighty participants were recruited through Prolific and were paid $9.50 for completing Experiment 2. This sample size was chosen based on the same power calculation as Experiment 1 and is enough participants to detect a medium effect size difference ($d = .5$) in final test performance among the errorless, errorful with immediate feedback, and errorful with delayed feedback conditions.

Three participants from the errorless+immediate condition were excluded for not following task instructions and providing a guess on at least 75% of the guess trials. Among the remaining 177 participants, sixty participants were randomly assigned to the errorless condition (mean age = 28.37 years, mean education = 14.54 years, 19 female, 40 male, 1 non-binary), 58 participants were randomly assigned to the errorful+immediate condition (mean age = 29.00 years, mean education = 15.41 years, 21 female, 35 male, 2 non-binary), and 59 participants were randomly assigned to the errorful+delayed condition (mean age = 28.97 years, mean education = 14.80 years, 27 female, 32 male). All participants were fluent English speakers.

### ***Materials, Design, and Procedure***

Participants were randomly assigned to one of three between-subjects conditions: errorless learning, errorful learning with immediate feedback (errorful+immediate), and errorful learning with delayed feedback (errorful+delayed). As in Experiment 1, participants learned 60 weakly related word pairs under errorful or errorless learning conditions, completed a 5-minute arithmetic distractor task, took a final test on the pairs, and then completed the study strategies inventory (Figure 11).

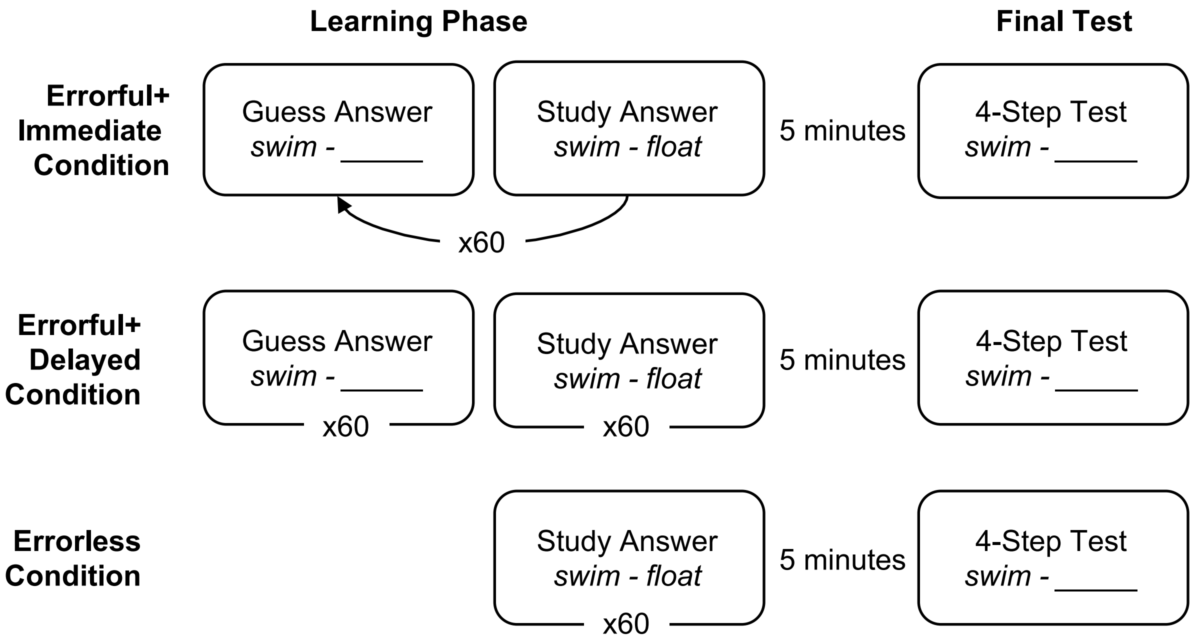


*Figure 11.* Experiment 2 design. Participants were randomly assigned to either the errorful+immediate, errorful+delayed, or errorless learning condition. Participants studied 60 weakly related word pairs, solved arithmetic problems as a 5-minute distractor task, and then took a cued recall test. However, in the errorful learning conditions, participants guessed the answer before studying the correct answer for each pair in the learning phase. Participants either

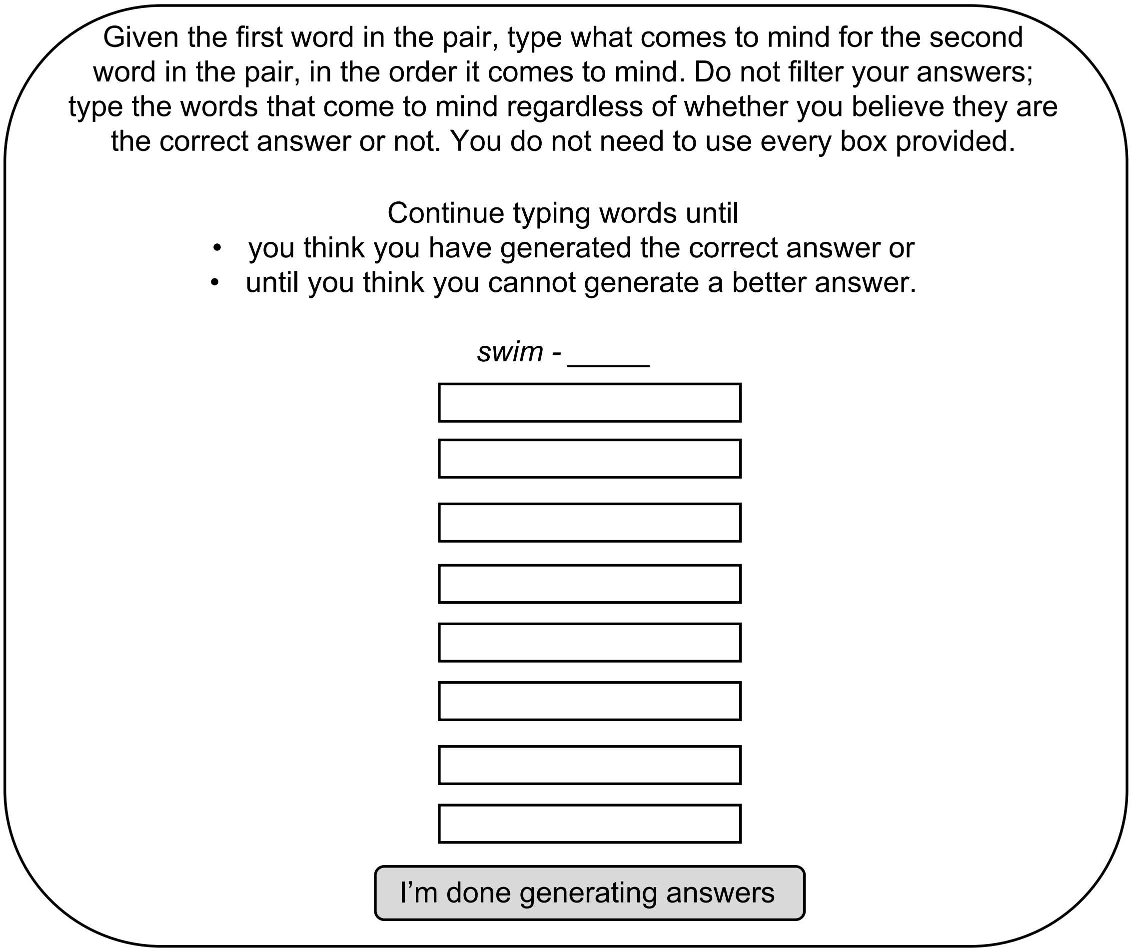
made a guess and then immediately learned the correct answer for each pair (errorful+immediate) or made a guess for all 60 pairs and then learned the correct answers for all 60 pairs (errorful+delayed).

During the learning phase, the errorless condition in Experiment 2 was nearly identical to the errorless condition in Experiment 1 and the errorful+immediate condition in Experiment 2 was nearly identical to the errorful condition in Experiment 1. However, unlike Experiment 1, counterbalanced item orders were not used. Instead, the pairs were presented in a random order for each participant during the learning phase and on the final test. Furthermore, time to study the correct answer was also controlled by the computer rather than the participant. Each pair was presented, and participants had 6 seconds to study the pair before the computer automatically advanced to the next trial.

A third between-subjects condition was added as well. The errorful+delayed condition was similar to the errorful+immediate condition. However, participants in the errorful+delayed condition made guesses for all 60 pairs before studying the correct answers for all 60 pairs. The pairs were presented in a different randomized order for guessing the answers, studying the correct answers, and the answer on the final test. As in the errorful+immediate condition, participants had as long as they needed to guess the answer for each pair and were then 6 seconds to study each correct answer.

A different final test format was used in Experiment 2 in order to isolate the effects of front-end and back-end retrieval processes (Halamish et al., 2012). Participants were instructed that the experiment was interested in “what goes through people’s minds when they try to remember something.” In the first step (Figure 12), the cue was presented and participants were instructed to write down every candidate target word that came to mind, in the order they came to mind, without any screening, regardless of whether they believed it was the correct answer.

Participants were instructed to stop searching memory and reporting candidate targets when they believed they had produced the correct answer or could no longer produce a better answer.



Given the first word in the pair, type what comes to mind for the second word in the pair, in the order it comes to mind. Do not filter your answers; type the words that come to mind regardless of whether you believe they are the correct answer or not. You do not need to use every box provided.

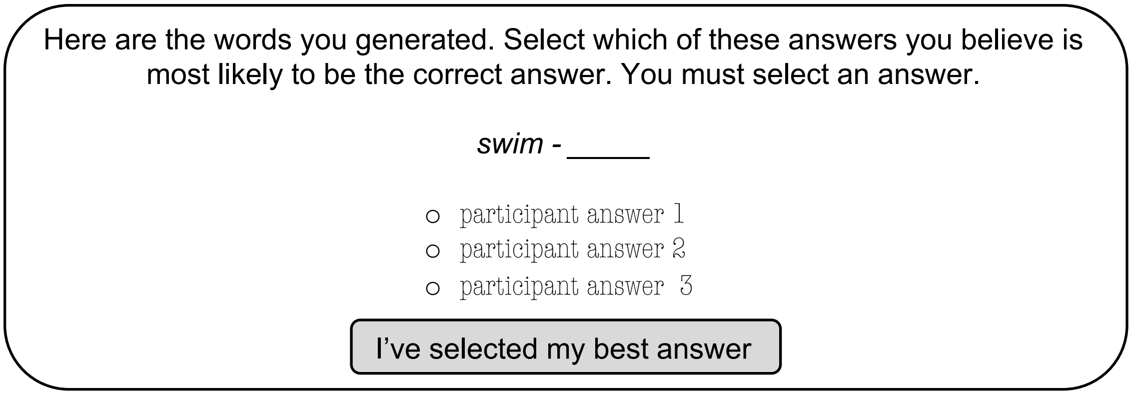
Continue typing words until

- you think you have generated the correct answer or
- until you think you cannot generate a better answer.

*swim* - _____

*Figure 12*. Step 1 of the four-step final test of Experiment 2. Participants reported all of the candidate answers that come to mind.

In step 2, if participants generated more than one candidate answer, they indicated the answer they believed most likely to be correct (Figure 13).



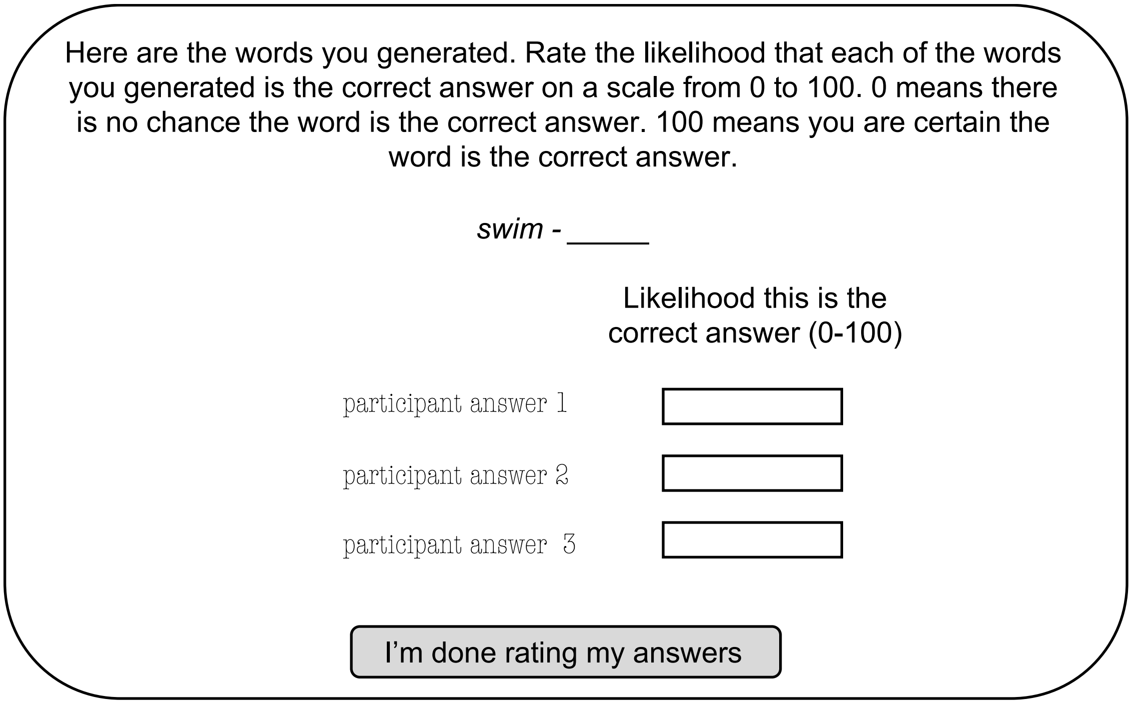
Here are the words you generated. Select which of these answers you believe is most likely to be the correct answer. You must select an answer.

*swim - _____*

- ○ participant answer 1
- ○ participant answer 2
- ○ participant answer 3

*Figure 13*. Step 2 of the four-step final test of Experiment 2. Participants selected the best answers among the candidate answers generated in step 1.

In step 3, participants rated the likelihood on a scale from 0-100 that each of the candidate answers they generated was correct (Figure 14).



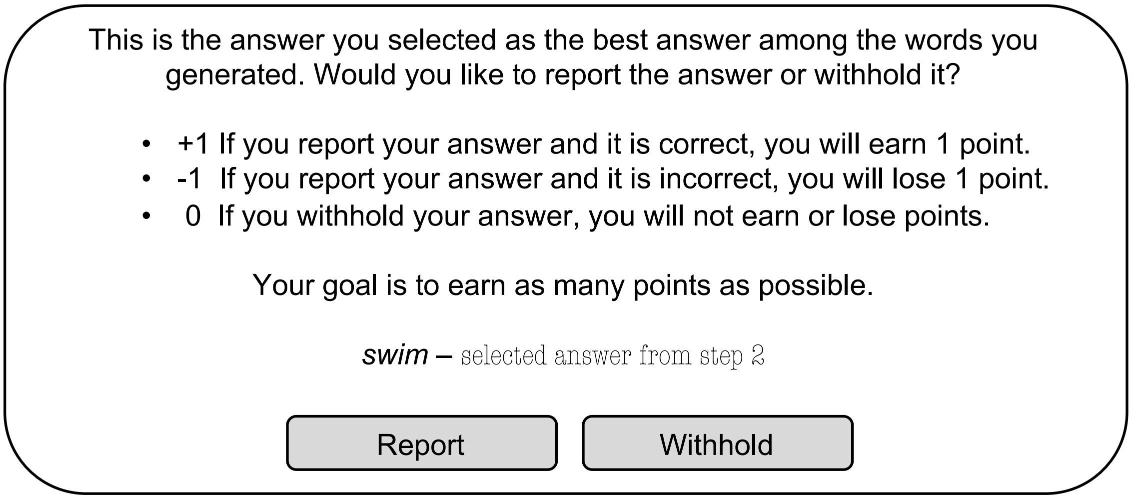
Here are the words you generated. Rate the likelihood that each of the words you generated is the correct answer on a scale from 0 to 100. 0 means there is no chance the word is the correct answer. 100 means you are certain the word is the correct answer.

*swim - _____*

| | Likelihood this is the correct answer (0-100) |
|---|---|
| participant answer 1 | |
| participant answer 2 | |
| participant answer 3 | |

*Figure 14*. Step 3 of the four-step final test of Experiment 2. Participants rated the likelihood that each candidate answer they generated in step 1 is the correct answer.

In the fourth and final step, participants chose whether to report or withhold the best candidate answer they selected in step two (Figure 15). One point was gained for correct answers reported; one point was lost for incorrect answers reported; no points were gained or lost for answers withheld. Participants were instructed that their goal was to earn as many points as possible and their current point total was displayed.



This is the answer you selected as the best answer among the words you generated. Would you like to report the answer or withhold it?

- +1 If you report your answer and it is correct, you will earn 1 point.
- -1 If you report your answer and it is incorrect, you will lose 1 point.
- 0 If you withhold your answer, you will not earn or lose points.

Your goal is to earn as many points as possible.

*swim* – selected answer from step 2

*Figure 15*. Step 4 of the four-step final test of Experiment 2. Participants decided whether to report or withhold the best candidate answer selected in step 2.

**Analysis Plan**

The purpose of Experiment 2 was to examine how errorful learning conditions affected memory in comparison to errorless learning as well as examine the effects of feedback timing on learning from errors. The final test format yielded multiple measures of memory, which are summarized in Table 3. In addition to measuring recall of the correct answer, these memory measures can differentiate front-end and back-end retrieval processes at the time of the final test.

Table 3

*Dependent Memory Measures in Experiment 2*

| Measure Name | Retrieval Process | Measure Description |
|---|---|---|
| forced-report accuracy | overall accuracy | number of correct answers selected as the best candidate answer on step 2 divided by the total number of trials |

|  |  |  |
|---|---|---|
| free-report accuracy | overall accuracy | number of correct answers selected as the best candidate answer on step 2 divided by the number of answers reported on step 4 |
| target accessibility | front-end | number of trials on which the correct answer was generated as one of the candidate answers divided by the total number of trials |
| number of candidates | front-end | average number of candidate responses generated across all trials |
| error-target co-generation | front-end | number of trials on which the error and correct answer were generated divided by the number of trials on which the correct answer was generated |
| Mediation | front-end | number of trials on which the error was generated as the first candidate, the correct answer was generated as the second candidate, and the correct answer was selected on step 2 divided by the total number of trials |
| confidence-accuracy resolution | back-end | gamma correlation between confidence in best candidate answer selected on step 2 and accuracy of that answer |
| reporting-accuracy resolution | back-end | gamma correlation between whether an answer is reported in step 4 and whether that answer is correct |
| accuracy improvement | back-end | difference between free-report and forced-report accuracy |

Given the four-step final test format of Experiment 2, memory for the correct answer can be measured in two ways. The better an individual's memory for the correct answer is, the more likely that the correct answer should be selected among all of the candidate answers on step 2 of the test. Because participants were required to select an answer on step 2, this measure of memory for the answer is referred to as *forced-report accuracy*. Similarly, the better an individual's memory for the correct answer is, the more likely that the correct answer will be reported on step 4 of the test. The proportion of answers that participants chose to report on step 4 of the test that were correct is referred to as *free-report accuracy*.

In order to report the correct answer on step 2 and step 4 of the final test, participants must retrieve the correct answer as one of the candidate answers on step 1 (a front-end retrieval process) and be able to select the correct answer among any other candidate answers that come to mind (a back-end retrieval process). *Target accessibility* indicates how frequently the target was generated as one of the candidates (regardless of whether it was selected as the correct answer). The *number of candidates* reported on step 1 of the test yields insights into front-end retrieval processes, specifically, whether the correct answer is typically retrieved alone or along with other candidate answers. If one's original error facilitates retrieval of the correct answer, then the error should typically be retrieved on trials on which the target is also retrieved. Therefore, *error-target cogeneration* should be significantly more likely than not. Similarly, if high levels of target recall are achieved, then *mediation* should occur on a high proportion of trials (i.e., the error is first retrieved, followed by the target).

If multiple candidate answers are reported on step 1, participants must engage in back-end retrieval processes, specifically post-retrieval monitoring, in order to determine which candidate answer is the correct answer (step 2). If participants are able to select the correct answer among multiple candidate answers, then participants must be able to differentiate between correct and incorrect answers. Participants who are able to select the correct answer must have relatively accurate insight into the likelihood that an answer they produced is correct (step 3). Therefore, the better the insight that participants have into the accuracy of an answer, the more their confidence ratings on step 3 should differentiate between correct and incorrect answers. That is, the confidence ratings provided on step 3 for the answer selected on step 2 should tend to be higher for correct answers than incorrect answers, as indicated by higher *confidence-accuracy resolution*.[2] Confidence-accuracy resolution was measured as the gamma correlation between confidence in best candidate answer selected on step 2 and accuracy of that answer. A stronger correlation, or higher resolution, indicates that confidence ratings better distinguished correct from incorrect answers.

Similarly, if participants have insight into the accuracy of an answer, then participants should tend to report the answers that are correct and withhold answers that are incorrect on step 4 of the test. That is, *reporting-accuracy resolution* should be high. Reporting-accuracy resolution was measured as the gamma correlation between whether that answer is reported in step 4 and whether that answer was correct. A stronger correlation, or higher resolution, indicates that participants' decisions to report or withhold an answer better reflect the accuracy of the answer.

The final measure of post-retrieval monitoring accuracy compares forced-report and free-report final test accuracy. On the forced-report portion of the final test (step 2), participants were required to provide an answer and therefore may provide an answer they know is incorrect. Therefore, forced-report final test accuracy conflates the ability to generate the correct answer

---

[2] Eight participants (3 from the errorful+immediate condition, 5 from the errorful+delayed condition) of the 177 participants were excluded from analyses involving confidence ratings. These eight participants did not provide confidence ratings on at least 75% of trials on step 3 on the final test. Additional participants had to be excluded from individual analyses involving gamma correlations when gamma was undefined for the participant. Therefore, the degrees of freedom may not align across analyses. See Appendix B for more details.

and the ability to identify whether an answer is correct, i.e., it conflates front-end and back-end retrieval processes. Free-report accuracy, in contrast, allows participants to only report a portion of their answers. A participant could theoretically achieve perfect free-report accuracy by recalling only one correct answer but reporting only that correct answer and none of their incorrect answers. Therefore, the difference between free-report and forced-report final test accuracy reflects one's ability to distinguish between correct and incorrect answers. If participants have insight into the accuracy of an answer, then participants will tend to choose to withhold their incorrect answers and report their correct answers from step 2, in which case, free-report accuracy will be higher than forced-report accuracy. That is, the accuracy improvement from the forced-report step 2 to the free-report step 3 provides a third measure of post-retrieval monitoring, i.e., the ability to differentiate between correct and incorrect answers (e.g., Koriat & Goldsmith, 1996).

These memory measures were used in Experiment 2 to reveal the mechanism by which making an error affects learning of the correct answer relative to errorless learning. These memory measures were also used to examine how delaying feedback effects learning from errors.

## Results

### *Guessing Accuracy*

Consistent with existing norms (Nelson et al., 1998), participants in the errorful conditions guessed the correct answer on 6% of trials; these trials were excluded from subsequent analyses.

### *Final Test Accuracy*

The format of the final test yielded two measures of memory accuracy: forced report accuracy and free report accuracy (Figure 16). When given the opportunity to report or withhold each answer, participant in the errorless, errorful+immediate, and errorful+delayed conditions reported 66%, 76%, and 69% of their answers (on average). The opportunity to withhold answers led to an increase in test performance. A 3 (study condition: errorless vs. errorful+immediate vs. errorful+delayed) x 2 (score type: forced report vs. free report) mixed-design ANOVA revealed a main effect of study condition, $F(2,177) = 13.14$, $p < .001$, $\eta_p^2 = .13$, such that errorless learning conditions led to poorer final test performance than errorful learning with immediate or delayed feedback. Furthermore, a main effect of score type emerged, $F(1,177) = 385.38$, $p < .001$, $\eta_p^2 = .68$, such that free report accuracy was higher than forced report accuracy. However, no statistically significant interaction emerged, $F(2,177) = 1.25$, $p = .29$, $\eta_p^2 = .01$.

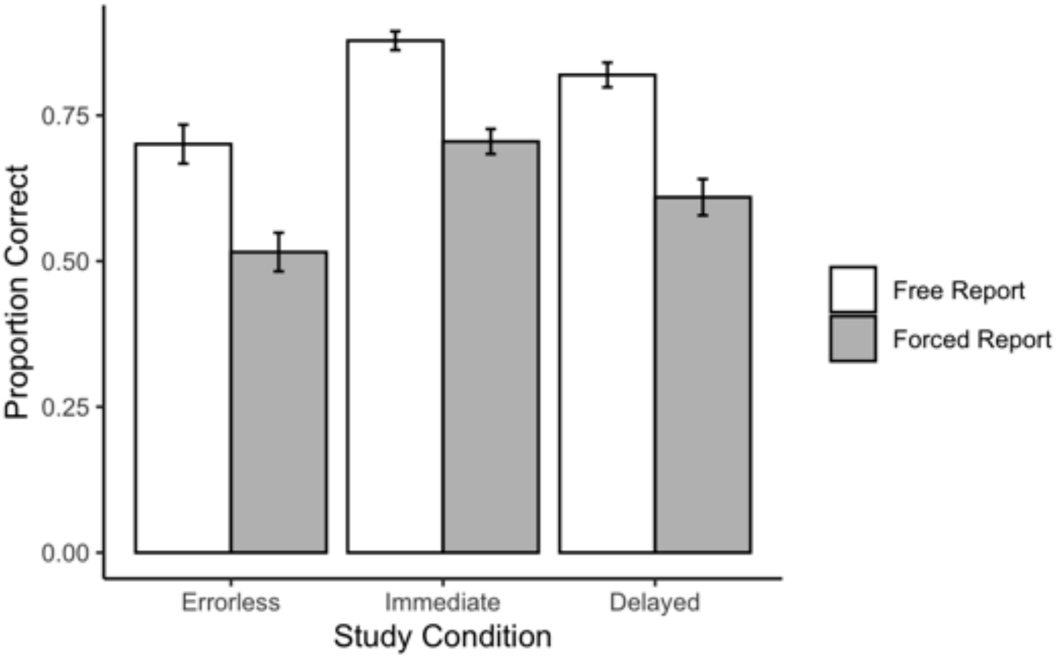


*Figure 16*. Proportion correct on the final test in Experiment 2. Forced report accuracy was calculated as the number of correct answers selected on step 2 among all trials. Free report accuracy was calculated as the number of correct answers reported on step 4 among the total number of answers the participant chose to report on step 4. Error bars represent ±1 standard error of the mean.

Planned comparisons revealed that errorful learning with immediate feedback led to significantly higher forced report accuracy than both errorful learning with delayed feedback, *t*(118) = 2.54, *p* = .01, *d* = 0.46, and errorless learning, *t*(119) = 4.82, *p* < .001, *d* = 0.88. The same pattern of results emerged for free report accuracy. Errorful learning with immediate feedback led to significantly higher free report accuracy than both errorful learning with delayed feedback, *t*(118) = 2.22, *p* = .03, *d* = 0.40, and errorless learning, *t*(119) = 4.81, *p* < .001, *d* = 0.87. Finally, forced report and free report accuracy were also significantly higher in the errorful+delayed than the errorless condition, *t*(117) = 2.07, *p* = .04, *d* = 0.38, *t*(117) = 2.99, *p* = .003, *d* = 0.55, respectively.

Thus, consistent with previous research, errorful learning enhanced memory relative to errorless learning, but errorful learning was more beneficial when immediate feedback was provided. Additional analyses examined measures of front-end and back-end retrieval processes to clarify the memory mechanisms underlying the benefits of errorful learning with immediate feedback over errorless learning and errorful learning with delayed feedback. Each comparison will be considered in turn.

### ***Errorful Learning with Immediate Feedback vs. Errorless Learning***

**Front-End Retrieval Processes.** Measures of front-end retrieval processes provide insight into how participants search memory and generate candidate answers. Participants in the errorless (*M* = 1.87, *SD* = 0.67) and errorful+immediate (*M* = 1.68, *SD* = 0.55) conditions generated a similar number of candidates on the final test, *t*(116) = 1.61, *p* = .11, *d* = 0.30. Thus, errorful learning with immediate feedback did not improve final test performance by reducing the number of candidate answers that were produced on the final test because even participants in the errorless condition generated a similar number of candidates. Instead, errorful learning with immediate feedback increased the probability that the target was one of the candidates produced. Participants in the erroful+immediate condition (*M* = .75, *SD* =.16) generated the correct answer as one of their candidate answers on step 1 of the final test on a significantly higher proportion of trials than participants in the errorless condition (*M* = .54, *SD* = .26), *t*(119) = 4.96, *p* < .001, *d* = 0.90. Thus, errorful learning made the correct answer more accessible on the final test than errorless learning.

However, other measures of front-end retrieval processes suggested that mediation cannot explain why the correct answer was more likely to be generated on the final test in the errorful+immediate condition than the errorless condition. Contrary to the errors-as-mediators hypothesis, error-target co-generation was low. Participants only generated their original error on 30% (*SD* = 17%) of the trials on which they also generated the correct target as one of the candidate answers. That is, when participants in the errorful+immediate condition generated the correct answer as one of their candidate answers, they were less likely than not to also generate their original error, *t*(57) = 9.12, *p* < .001, *d* = 1.20. Similarly, on only a small portion of trials did the pattern of participants' candidate generation reflect mediation, with the error generated followed by the target (*M* = 0.11, *SD* = 0.11).

In short, measures of front-end retrieval processes suggest that errorful learning with immediate feedback enhanced memory relative to errorless learning, in part, by increasing access to the correct answer on the final test, but not via recall of one's original error.

**Back-End Retrieval Processes.** There are two factors that influence accurate memory performance on the final test that are not mutually exclusive: generating the correct answer and selecting it among the other candidate answers on the final test. Front-end measures of retrieval processes suggested that errorful learning with immediate feedback increased the likelihood that participants generated the correct answer relative to errorless learning. Measures of back-end retrieval processes provide insight into participants' post-retrieval monitoring accuracy, i.e. their ability to evaluate the accuracy of an answer and select the correct one among the candidate answers.

As the analyses of free and forced report final test accuracy reported above revealed, post-retrieval monitoring was high, overall. Free-report accuracy was significantly higher than forced-report accuracy across all three conditions and the effect size was large ($d = 2.92$; Figure 16). However, there was no interaction effect, which suggests that participants in both the errorful+immediate and errorless conditions were similarly able to differentiate between correct and incorrect answers.

Other measures provided converging evidence for high post-retrieval monitoring accuracy, but also revealed differences between the errorful+immediate and errorless conditions. Confidence-accuracy resolution was strong overall but was significantly higher in the errorful+immediate condition ($M = 0.88$, $SD = .10$) than the errorless condition ($M = 0.77$, $SD = 0.22$), $t(108) = 3.36$, $p = .001$, $d = 0.64$.[3] Similarly, accuracy-reporting resolution was near 1, its maximum value, suggesting that participants were well able to distinguish between correct and incorrect answers. However, accuracy-reporting resolution was significantly higher in the errorful+immediate ($M = .97$, $SD = .06$) than errorless condition ($M = .89$, $SD = .19$), $t(109) = 3.07$, $p = .003$, $d = 0.58$.

Collectively, the measures of front-end and back-end memory processes suggest that errorful learning with immediate feedback not only enhanced the accessibility of the correct

---

[3] Descriptive statistics and additional analyses of confidence ratings are reported in Appendix C for completeness.

answer relative to errorless learning, but increased participants' ability to differentiate between correct and incorrect answers on the final test. However, there was no evidence for mediation, i.e., that participants recalled correct answers via their original errors.

### *Errorful Learning with Immediate vs. Delayed Feedback*

**Front-End Retrieval Processes.** The analyses of final test accuracy reported above were consistent with previous research: errorful learning conditions enhanced memory more when feedback was provided immediately after an error was made. However, the difference in final test accuracy cannot be attributed to the number of candidates generated. Participants in the errorful+immediate ($M = 1.68$, $SD = 0.56$) and errorful+delayed ($M = 1.65$, $SD = 0.58$) conditions generated few candidates and generated a similar number of candidates, $t(115) = 0.44$, $p = 0.66$, $d = 0.08$. Instead, the timing of feedback affected whether the correct target was generated at all. Participants in the erroful+immediate condition ($M = 0.74$, $SD = 0.15$) were significantly more likely than participants in the errorful+delayed condition ($M = 0.63$, $SD = 0.24$) to generate the correct answers as one of their candidate answers, $t(115) = 2.86$, $p = .005$, $d = 0.53$.

Again, there was no evidence that one's original error served as a mediator to retrieve the correct answer. Just as in the errorful+immediate condition, participants in the errorful+delayed condition only generated their original error on 25% ($SD = 16\%$) of the trials on which they also generated the correct target as one of the candidate answers. That is, when participants in the errorful+delayed condition generated the correct answer as one of their candidate answers, they were less likely than not to also generate their original error, $t(59) = 12.32$, $p < .001$, $d = 1.60$. Similarly, on only a small portion of trials did the pattern of participants' candidate generation reflect mediation, with retrieval of the error followed by retrieval of the target (*M* = 0.08, *SD* = 0.09).

Furthermore, differences in mediation cannot account for differences in final test performance. Although participants in the errorful+immediate condition recalled significantly more correct answers, participants in the two errorful conditions were similarly likely to also generate their error when they generated the target as one of their candidate answers, *t*(115) = 1.66, *p* = 0.10, *d* = 0.31. The participants in both errorful conditions were also similarly likely to show a pattern of mediation in their candidate generation, *t*(115) = 1.50, *p* = 0.14, *d* = 0.28. Thus, the errors-as-mediators hypothesis likely cannot account for why immediate feedback enhances learning from errors. Instead, the measures of front-end retrieval processes suggest that immediately presenting the correct answer after making an error enhances encoding of the answer, making it more accessible on the final test.

**Back-End Retrieval Processes.** Although delaying feedback after errors significantly impaired final test performance, it had only a slight negative effect on post-retrieval monitoring. As in the errorless and errorful+immediate conditions, participants in the errorful+delayed condition effectively distinguished between correct and incorrect answers. Free-report test accuracy was significantly higher than forced-report accuracy in all conditions, including in the errorful+delayed condition (Figure 16). Furthermore, both confidence-accuracy resolution (*M* = 0.79, *SD* = 0.30) and accuracy-reporting resolution (*M* =.92, *SD* = 0.27) were excellent in the errorful+delayed condition. Although confidence-accuracy resolution was significantly lower in the errorful+delayed condition than the errorful+immediate condition, *t*(106) = 2.08, *p* = .04, *d* = 0.40, no difference in accuracy-reporting resolution emerged, *t*(111) = 1.38, *p* = .17, *d* = 0.26.

Taken together, delaying feedback impaired participants' ability to distinguish between correct and incorrect responses only slightly, if at all.

### ***Self-Reported Study Strategies***

As in Experiment 1, Experiment 2 explored whether study condition affected participants' self-reported study strategies. The pattern of results was highly similar (Table 4, reported in order of frequency of use). Participants reported using both effective (e.g., mental imagery) and ineffective strategies (e.g., repetition). However, unlike Experiment 1, some statistically significant differences in strategy use across conditions emerged. Participants in the errorless condition were more likely to report using repetition to memorize the word pairs than participants in the errorful+immediate, $z = 3.60$, $p < .001$, or errorful+delayed conditions, $z = 2.14$, $p = .03$.

There were no differences in self-reported strategy use between the immediate and delayed errorful conditions, except that participants in the immediate condition were more likely to report using mental imagery than in the delayed condition, $z = 2.31$, $p = .02$. In addition, seventeen participants (approximately 10% of participants; 4 in the errorless condition, 5 in the errorful+immediate condition, and 8 in the errorful+delayed condition) selected 'Other' on the study strategies survey and described their own strategy that they had used. Most of the strategies that participants reported could be classified as one of the listed study strategies on the survey (e.g., "repeated the pair over and over by speaking it out loud" could be classified as repetition or "try to recall the less obvious ones now and again" could be classified as retrieval). Two of the responses that participants wrote in suggested some sort of mediation strategy to recall the target (e.g., "I also recalled a few of my first guesses" or "compared it to the word I had chosen earlier").

Table 4

*Proportion of Participants Reporting Use of Study Strategies in Experiment 1*

| Study Strategy | Errorless | Immediate | Delayed |
|---|---|---|---|
| Repeated the pair over and over again in your head | $0.70^{\dagger *}$ | $0.36^{\dagger}$ | $0.51^{*}$ |
| Made a mental image of the words in the pair | 0.57 | $0.69^{x}$ | $0.49^{x}$ |
| Related the word pair to yourself | 0.32 | 0.24 | 0.25 |
| Came up with a sentence or a story to connect the words in the pair | 0.23 | 0.21 | 0.15 |
| Acted out or imagined acting out the words in the pair | 0.17 | 0.16 | 0.15 |
| Tested yourself by covering up the second word in the pair | 0.03 | 0.07 | 0.10 |
| Created mnemonics such as rhymes or acronyms to connect the words in the pair | 0.07 | 0.05 | 0.03 |

*Note.* Superscripts indicate a statistically significant difference with $p < 0.05$ between proportions with the same symbol.

## Discussion

Consistent with prior research, errorful learning with immediate feedback enhanced memory. Both forced-report and free-report final test accuracy were higher in the errorful+immediate condition than in the errorless and errorful+delayed conditions. The purpose of Experiment 2 was to understand the memory mechanisms underlying the benefits of errorful learning with immediate feedback. Experiment 2 compared two different accounts of learning from errors: SEED and errors-as-mediators.

According to SEED, making an error enhances subsequent encoding of the correct answer, strengthening the cue-target association through strategic adaptations in time, attention, or qualitative study strategies. Therefore, SEED proposes that errorful learning with immediate feedback enhances memory for the correct answer on the final test through two non-mutually exclusive processes, corresponding to front-end and back-end retrieval processes, respectively. First, enhanced encoding of the target could improve the likelihood that the correct answer is retrieved by narrowing the search set on the final test to include the correct answer and exclude one's original error and other related words and concepts that could cause interference. Second, enhanced encoding of the target could increase recollection for the study episode, thereby increasing the likelihood that the correct answer is selected among candidate answers.

Consistent with the hypothesis that errorful learning with immediate feedback enhances the cue-target association, participants in the errorful+immediate condition were significantly more likely to generate the target as one of their candidate answers than participants in the errorless condition. Although errorful learning with immediate feedback did not reduce the number of candidates that were generated on the final test relative to errorless learning, participants in both conditions generated only one to two candidates, on average. Therefore, Experiment 2 provides novel evidence that the reason errorless learning is less effective is not due to the amount of information that comes to mind during a retrieval attempt, but rather, whether the correct answer is among the information that comes to mind. Errorful learning with immediate feedback appears to enhance encoding of the correct answer such that the target becomes more accessible and can be generated on the final test.

According to SEED, delaying feedback impairs learning from errors because the time and items intervening between when an error is made and when the correct answer is presented interferes with strategic encoding of the answer. That is, making an error is thought to provide participants with information regarding how difficult the item is for them to learn and how they should think about the item; the delay may make it difficult for participants to maintain this information and adapt their study time, attention, or study strategies for the correct answer accordingly (Carpenter & Vul, 2011). Consistent with the account that delaying feedback impairs subsequent encoding, participants in the errorful+delayed condition were significantly less likely to generate the target as one of their candidate answers than participants in the errorful+immediate condition.

The errors-as-mediators hypothesis provides a different account for the benefits of errorful+immediate learning, suggesting that errorful learning enhances memory by including one's original error in the memory search, not by excluding it. The cue is thought to bring to mind one's original error, which activates the target in turn, providing an additional retrieval route through which the target can be recalled. There was no evidence, though, that mediation could explain why participants were more likely to generate the target on the final test in the errorful+immediate condition than the errorless condition. When participants generated the correct answer, they only generated their original error as another candidate on a small subset of trials. Therefore, although Experiment 2 revealed that errorful learning with immediate feedback enhanced encoding of the correct answer, the evidence does support the conclusion that such enhanced encoding involved forming cue→error→target mediated pathways.

According to the errors-as-mediators hypothesis, delaying feedback impairs learning from errors because it interferes with forming the error-target association. Therefore, on the final test, the cue would bring to mind one's original error, but the error would be less likely to activate the target in turn (Bridger & Mecklinger, 2014; Cyr & Anderson, 2015, 2018; Kornell et al., 2009; Hays et al., 2013; Knight et al., 2012; Kornell, 2014; Vaughn & Rawson, 2012). However, contrary to this account, evidence of mediation was similarly infrequent in both the errorful+immediate and errorful+delayed conditions, even though memory accuracy was significantly worse in the errorful+delayed condition.

In short, prior research has used indirect methods (e.g., manipulating word pair relatedness) to infer that mediation plays a central role in learning from errors (e.g., Cyr & Anderson, 2015; Kornell et al., 2009). Experiment 2 offered a more direct test of the errors-as-mediators hypothesis and found little evidence for mediation as an explanation for why errorful learning with immediate feedback enhances memory relative to errorless learning or why errorful learning enhances memory more when feedback is provided immediately rather than at a delay (c.f., Vaughn & Rawson, 2012).

Although the design of Experiment 2 does not offer a definite alternative mechanism for how errorful learning enhances subsequent encoding of the correct answer, the measures of back-end retrieval processes offer some hints that are consistent with SEED. Specifically, enhanced encoding of the target in the errorful+immediate condition may have involved encoding of not just the semantic association between the cue and the target, but other details of the study episode as well. On the final test, when participants use the cue to recall the answer, specific details of the study episode may also come to mind. Participants can use the quantity and quality of the study episode details recollected to evaluate whether the answer that they generated is likely correct. For example, if participants recollect details of the study experience (e.g., the feeling of surprise that their guess was wrong, the other information they were thinking about at the time, a noise in the hallway at the time the target was presented, etc.) then participants can feel confident that the answer they generated alongside those episodic details is likely correct. In contrast, if few episodic details are recollected with the answer they generated, then participants may feel less confident that their answer was actually presented in the experiment and is correct (Kelley & Sahakyan, 2003).

Consistent with this episodic discrimination explanation, errorful learning with immediate feedback improved the accuracy of back-end retrieval processes. Participants in all three conditions were able to effectively determine whether the answer they had selected on step 2 was correct. However, participants in the errorful+immediate condition showed even better confidence-accuracy resolution and reporting-accuracy resolution than participants in the errorless condition. The fact that participants were better able to distinguish between correct and incorrect answers in the errorful+immediate than errorless condition suggests that errorful learning may have enabled participants to encode not only the semantic association between the cue and the target, but also other details of the study episode.

However, there is no clear explanation for why there were small to negligible differences in post-retrieval monitoring between the two errorful conditions, even though target accessibility was significantly lower in the errorful+delayed condition than the errorful+immediate condition. If delaying feedback impaired strategic encoding of the correct answer, then delaying feedback should have also impaired encoding of details of the study episode as well and differences in post-retrieval monitoring should have emerged. Perhaps it was difficult to observe difference in accuracy-reporting resolution between the two errorful conditions because resolution was near ceiling in both conditions. Future research should examine the degree to which the quality of initial encoding and post-retrieval monitoring accuracy are associated (e.g., Kelley & Sahakyan, 2003).

In sum, Experiment 2 revealed that errorful learning enhanced memory for the correct answer relative to errorless learning, particularly if feedback was provided immediately after the error was made. The results of Experiment 2 were more consistent with SEED than the errors-as-mediators account of the mechanisms underlying learning from errors.

# CHAPTER 4 – GENERAL DISCUSSION

No one wants to make a mistake when performance matters, whether it be in a championship game, on a final exam, or in a new job. However, errors are an inevitable part of learning something new. It is nearly impossible to learn how to speak Norwegian fluently or how to calculate complex derivatives without making mistakes along the way. An important practical and theoretical question is how making mistake while learning affects performance in the long-term. One concern is that committing an error will only reinforce the mistake in memory, making it more likely to be repeated again in the future. From this perspective, mistakes should be avoided while acquiring a new skill or knowledge. Despite this concern, research has revealed that making mistakes can be a potent learning opportunity for encoding new, correct information (for a review, see Metcalfe, 2017). Consistent with prior research, the present studies revealed that making errors while learning enhanced memory for the correct answer relative to errorless learning (Experiments 1 and 2; $d = 0.78$), and the benefits of errorful learning were larger when feedback was immediately provided, rather than after a delay of several minutes (Experiment 2). The purpose of the present studies was to test theoretical mechanisms by which errorful learning enhances memory. Experiments 1 and 2 tested predictions of the newly proposed SEED model; Experiment 2 compared SEED to the established errors-as-mediators account of how generating errors enhances learning.

## Errors-as-Mediators

The errors-as-mediators hypothesis is a leading explanation for why making an error would be beneficial for learning. The errors-as-mediators hypothesis proposes that errors serve as “stepping stones” from the question to the answer, enhancing memory for the correct answer by providing an additional retrieval route (e.g., Cyr & Anderson, 2015; Kornell et al., 2009). That is, recalling one's previous mistake is thought to be a critical step in recalling the correct answer. A mediator-based account has also been put forth for why delaying feedback after making an error reduces the benefits of errorful learning conditions (Hays et al., 2013; Kornell, 2014; Vaughn & Rawson, 2012). That is, if feedback is delayed, then one's error and the correct answer are not activated simultaneously, thereby decreasing the likelihood that the error and target become associated in memory. Therefore, on the final test, the cue may bring to mind one's original error, but the error is less likely to activate the correct answer, in turn.

However, the existing evidence for the errors-as-mediators hypothesis is indirect. Previous research has manipulated material type under the assumption that semantically richer materials (e.g., related word pairs, general knowledge trivia questions) afford more opportunity for the association between the error and answer to form than more semantically impoverished materials (e.g., unrelated word pairs, word stems). Indeed, generating errors has been shown to enhance memory for the correct answer for semantically richer materials, but not more semantically impoverished materials (e.g., Cyr & Anderson, 2015; Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012; Kornell, 2014). Other differences between materials or other mechanisms could explain why generating errors enhances memory for semantically rich, but not semantically impoverished materials.

Experiment 2 provided a novel, more direct test of the errors-as-mediators account of error correction. Rather than manipulating the type of materials to infer that errors serve as mediators, Experiment 2 used the four-step final test to examine the memory processes involved with retrieving the correct answer on the test. On step 1, participants were asked to generate all of the answers that came to mind in the order that they came to mind. If participants recalled their error and then used their error to retrieve the correct answer, then participants in the errorful+immediate condition who achieved high levels of memory accuracy (71% forced report accuracy) should frequently generate their error as a candidate answer, particularly as the first candidate generated. Contrary to this prediction, participants in Experiment 2 only generated their error on approximately one third of trials (37%) and generated it as their first candidate answer on only approximately one quarter of trials (28%). As further evidence against the hypothesis that errors facilitate retrieval of the correct answer, participants in the errorful+immediate condition were more likely to generate the correct answer alone than generate their error and the correct answer. Thus, errorful learning with immediate feedback produced high levels of memory for the correct answer, but there was no evidence that errors mediated retrieval of the correct answer.

The errors-as-mediators hypothesis also cannot account for why delaying feedback interfered with the benefits of learning from errors. Memory accuracy was lower in the errorful+delayed condition than the errorful+immediate condition, even though the proportion of trials on which participants demonstrated mediation—i.e., generated the error first, the correct answer second, and then selected the correct answer on step 2—was similar in the two conditions, albeit low (10% of trials or fewer).

One possible limitation of Experiment 2 is that errorful learning with immediate feedback enhanced memory for the correct answer via mediated cue→error→target pathways but the four-step final test format was not effective for capturing this retrieval process. Perhaps participants filtered the candidate answers that came to mind and did not report their error when it came to mind first because they knew it was incorrect. Although an example was provided at the start of the final test and instructions were provided on every trial to not filter the candidates that came to mind, it remains possible that participants recalled their errors initially but did not report them. However, previous research has successfully used this four-step final test format to reveal how different encoding conditions affect the quantity of information that comes to mind during a retrieval attempt (Halamish et al., 2012; Thomas & McDaniel, 2013). Therefore, the four-step final test is likely capable of detecting whether participants in the errorful+immediate and errorful+delayed conditions frequently generated their error as one of the candidates on step 1 of the final test.

Nevertheless, future research could use alternative methods for examining whether errors mediate correct answer retrieval on the final test. For example, in order to examine the strength of the error-target association, the final cued-recall test could provide participants' errors rather than the original cue and instruct participants to recall the target (Carpenter, 2011; Cho, Neely, Brennan, Vitrano, & Crocco, 2017; Coppens, Verkoeijen, Bouwmeester, & Rikers, 2016; Rawson, Vaughn, & Carpenter, 2015; Vaughn & Rawson, 2012).

### SEED

Experiment 2 revealed that participants in the errorful+immediate and errorful+delayed conditions typically retrieved the correct answer as the only candidate that came to mind on step 1 of the final test and did not generate their original error first. This pattern of results does not support the errors-as-mediators hypothesis and, instead, is consistent with the principle that memory tends to be best when the cue is uniquely associated with the target, rather than the cue being associated with multiple pieces of information, (e.g., the error and answer; Moscovtich & Craik, 1976; Nairne, 2002; Raaijmakers, 2003; Raaijmakers & Shiffrin, 1981; Watkins & Watkins, 1975). The issue of cue uniqueness has been forwarded as a critique of the errors-as-mediators hypothesis (Grimaldi & Karpicke, 2012). The present studies proposed and tested

SEED, a new explanation for why making errors would be beneficial for learning that is consistent with cue uniqueness.

The strategic encoding (SE) component of SEED suggests that making an error enables individuals to effectively adapt how they encode each correct answer, thereby selectively strengthening the question-answer association. Some of the ways in which participants could strategically adapt their encoding to learn from errors could include increasing attention (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009; Peterson & Wissman, 2020), allocating more study time to more difficult items (Soderstrom & Bjork, 2014), and/or qualitatively shifting encoding strategies to meet their study needs for each item (e.g., Pyc & Rawson, 2010; deWinstanley & Bjork, 2004).

SEED suggests that strategic encoding enhances later memory performance through front-end and/or back-end retrieval processes (for reviews, see Goldsmith & Koriat, 2007; Koriat et al., 2008). According to SEED, strategic encoding affects front-end retrieval processes on the final test such that the search set is narrowed and only the correct answer comes to mind at the exclusion of one's original error. The episodic discrimination (ED) component of SEED refers to how strategic encoding during the initial learning phase could enhance final test performance during back-end retrieval processes. According to SEED, generating an error enhances encoding of not only the correct answer, but details of the study episode. Participants can later use these episodic memory details to differentiate the correct answer from the other candidate answers that come to mind during the retrieval attempt.

Experiment 1 tested the strategic encoding hypothesis component of SEED. Specifically, Experiment 1 tested the hypothesis that making an error provides participants with idiosyncratic information regarding how difficult each item is for them to learn and therefore effectively adapt how long they study each item (Soderstrom & Bjork, 2014). If errorful learning enhanced memory by engendering strategic adaptations in study time, then a participant's item-by-item test performance should be better predicted by their own study times rather than another randomly-selected participant's study times and this should be true more so in the errorful than the errorless condition. In contrast to this prediction of SEED, personal study times did not predict recall on the final test better than another participant's study time in the errorful condition. In fact, in contrast to previous research, in both conditions, study times predicted recall on the final test only weakly, if at all.

One explanation for the weak study time-recall association has to do with how the materials were constructed and how the study condition was manipulated. In Experiment 1, the cue-target pairs had almost identical association strengths meaning that the pairs were all normatively the same difficulty to learn. Perhaps the pairs were too homogenous and gave participants nearly the same subjective experience of difficulty on every trial and therefore did not yield enough intra-individual variability in study times. For example, in the errorful condition, participants' guesses were wrong on nearly every trial. Therefore, after the first few trials, participants were likely not surprised that their guesses were wrong, making the subjective difficulty of each trial relatively similar (i.e., a wild guess followed by studying a different answer). Indeed, previous research that has demonstrated how study times can reflect strategic encoding have utilized more heterogenous learning materials (e.g., participants learned a mix of related and unrelated pairs; Koriat et al., 2006; Soderstrom & Bjork, 2014). Using a similar approach as Experiment 1, future research could examine whether errorful learning engenders more strategic allocation of study time than errorless learning when there is more variability in the learning experience (e.g., participants could study related and unrelated pairs, a portion of participants' guesses could be deemed correct, study condition could be manipulated within-subjects, etc.).

Regardless of the explanation for why study times were not predictive of recall in Experiment 1, the conclusion is the same: large differences in test performance emerged between the errorful and errorless conditions, but study time cannot explain these memory differences. Experiment 2 used the four-step final test format (Goldsmith & Koriat, 2007; Halamish et al., 2012; Thomas & McDaniel, 2013) to investigate how higher final test accuracy was achieved following errorful than errorless learning. Experiment 2 revealed that one of the ways errorful learning with immediate feedback enhanced final test performance relative to errorless learning was through front-end retrieval processes, namely, by increasing the likelihood that the correct answer was generated as one of the candidate answers on the test. The fact that the target was more accessible on the final test in the errorful condition than the errorless condition is consistent with SEED and suggests that generating an error enhanced encoding of the correct answer once it was presented, which strengthened the cue-target association.

Experiment 2 provided preliminary evidence for what the nature of such enhanced encoding after an error could reflect, besides changes in study time. Not only were participants in the errorful+immediate condition more likely to generate the correct answer on the final test than participants in the errorless condition, but they were better able to evaluate the accuracy of their answer. Participants' confidence ratings better distinguished between correct and incorrect answers in the errorful+immediate condition than the errorless condition. Similarly, participants' decisions to report or withhold an answer on step 4 of the final test more closely aligned with whether that answer was correct in the errorful+immediate condition than the errorless condition.

The fact that post-retrieval monitoring was higher in the errorful+immediate condition can be explained by participants using episodic memory to discriminate between correct and incorrect answers. Previous research suggests that making an error could improve recollection of the details of the initial study phase on the final test (e.g., Bishara & Jacoby, 2008; Chan & McDermott, 2007) and that participants could use these recollected details from the study phase to evaluate the accuracy of the answers that they generated (Kelley & Sahakyan, 2003) in order to produce the correct answer rather than other competitive alternatives (Anderson & Craik, 2006). Thus, although the post-retrieval monitoring results are consistent with SEED, future research should gather more direct evidence for the role of episodic memory in learning from errors.

In sum, Experiment 2 revealed that errorful learning, at least with immediate feedback, improved the accuracy of both front-end and back-end memory process relative to errorless learning. Specifically, higher final test performance in the errorful+immediate than in the errorless condition in Experiments 1 and 2 can be attributed to the fact that errorful learning enhanced not only the likelihood that the correct answer was generated on the final test (reflecting strategic encoding), but also the participants' ability to identify it as the correct answer (reflecting the use of episodic memory to discriminate between correct and incorrect information). However, there was no evidence that errors served as mediators to help participants retrieve the correct answer. Thus, the evidence from the present studies was more consistent with SEED than the errors-as-mediators hypothesis.

It remains an open question whether enhanced encoding of the correct answer after an error is indeed a strategic (i.e., controlled) process as SEED suggests. According to SEED, enhanced memory for the correct answer after making an error reflects intentional shifts in encoding effort and/or strategies. An alternative explanation is that making an error enhances subsequent encoding of the correct answer through automatic processes. One possible automatic process is retrieval-induced memory malleability. That is, retrieving previously learned information can make memories traces more malleable and more likely to be updated with new, relevant information (for a review, see Finn, 2017). There is evidence that retrieval-induced memory updating can happen automatically. For example, research has shown that retrieving previously learned information can lead to effective incidental encoding of newly presented information. Buckner and colleagues (2001) had participants learn a list of words and then take a recognition test involving old words and new foils in which participants indicated whether each word had been studied initially. In the intentional learning condition, participants were also instructed to memorize all of the words on the first recognition test, including the new lures. In the incidental learning condition, participants were given no instructions about memorizing the recognition test words. On a surprise second recognition test, the initial list of words was presented along with the foils from the first recognition test and new foils; participants indicated whether each word had been presented on the first recognition test. On this second recognition test, memory for the foils from the first recognition test was equivalent in the intentional and incidental conditions. Thus, retrieving old information can enable effective automatic encoding of new information that is simultaneously presented (see also Jacoby et al., 2005).

Through a similar automatic process, generating an erroneous answer may make the memory trace more malleable and thus more likely to be updated with the correct answer. However, if generating an error automatically makes the memory trace more likely to be updated with the correct answer, then generating an error should always enhance encoding of the correct answer once it is presented. In contrast to this prediction, generating errors does not enhance learning of semantically impoverished materials, such as unrelated word pairs or word stems (Bridger & Mecklinger, 2014; Cyr & Anderson, 2015; Grimaldi & Karpicke, 2012; Huesler & Metcalfe, 2012; Knight et al., 2012; Kornell, 2014). Therefore, the materials-specific benefits errorful learning suggests that error correction is not an automatic process.

## Future Research

Future research should further examine the likely controlled processes that make encoding of the correct answer more effective following an error than under errorless learning conditions. Specifically, future research should aim to collect more direct evidence that generating an error leads to enhanced strategic encoding of details of the study episode relative to errorless learning. One approach would be to have participants learn information under errorful or errorless conditions as in the present studies. However, on the final test, the target participants would be tested on their memory for the correct answer, but also details associated with the study episode. For example, the computer screen could be divided into quadrants and each pair could appear in a randomly selected quadrant for studying. On the final test, participants would be presented tested on not only their memory for the correct answer but details of the study episode, i.e., where the pair had been presented on the screen for initial study (for a similar approach, see Akan et al., 2018). Memory for study quadrant should be higher following errorful learning than errorless learning.

Additional studies should examine the role of attention in learning from errors by manipulating attention during encoding. If making an error increases attention, which enhances encoding of the correct answer and associated study details (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009; Peterson & Wissman, 2020), then the negative effects of divided attention on learning from errors should be evident on a four-step final test like the one used in Experiment 2.

Specifically, dividing attention during encoding of the correct answer should reduce the likelihood that the correct answer will be generated on the final test in both errorful and errorless conditions, but divided attention should have a larger negative impact in the errorful condition. Furthermore, if participants use recollected details of the study episode to evaluate the accuracy of their generated answers, then divided attention at encoding should impair post-retrieval monitoring accuracy more in an errorful than errorless condition (Kelley & Sahakyan, 2003).

## Practical Implications

The present studies were theoretically driven with the goal of understanding how generating errors enhances memory for correct answers. However, they have practical implications as well. One assumption of SEED is that the experience of making a mistake provides the learner with idiosyncratic information regarding how difficult the correct answer will be for them to learn so that they can adapt their encoding accordingly (e.g., Soderstrom & Bjork, 2014). One consequence of this assumption is that it is essential for learners to make their own errors, not merely learn what common mistakes other people have made. Indeed, Metcalfe and Xu (2018) found that errorful learning led to better memory for correct answers than errorless learning, but only when errorful learning involved generating one's own error, not reading the error another participant had made. Therefore, in classrooms, the instructor should encourage students to think of or write down their own answer to each question, even if they are not sure what the answer is or do not want to volunteer to answer the question out loud. SEED suggests that this experience of failing to come up with an answer or generating an incorrect answer will help students more effectively encode the correct answer when the instructor explains it and remember that moment in more detail on a later test.

In addition, Experiment 2 revealed that the benefits of making errors diminish when feedback is delayed (c.f., Kornell, 2014). In a real classroom or other learning environment, feedback often cannot come immediately after each question. For example, students in a Biology course may complete a homework assignment before it is due on Friday, but not get feedback on their work until the following Friday. Experiment 2 revealed that memory was worse in the errorful+delayed condition than the errorful+immediate condition because the target was less accessible, suggesting that delaying feedback impaired participants' ability to effectively encode the correct answer once it was presented. Perhaps in a practical setting, instructors could implement a "feedback review" assignment in which students are encouraged to do more than just read the feedback. For example, students could be instructed to explain their mistakes and why the correct answers are correct (e.g., Woloshyn, Pressley, & Schneider, 1992). Such elaborative processing of the feedback may encourage deeper encoding and support learning.

Finally, the present studies tested SEED with simple materials—i.e., related word pairs. Nevertheless, SEED has implications for more realistic and complex learning materials, such as calculating an integral, explaining the difference between force and work, or drawing a diagram of photosynthesis. The strategic encoding component of SEED posits that making an error enhances memory by enabling individuals to effectively adapt how they encode the correct answer. Therefore, a consequence of SEED is that in order to benefit from making errors, one must have knowledge of what these effective strategies could be. As a result, making an error may not enhance learning of correct information in an unfamiliar domain. For example, consider a student who knows little about biology and erroneously guesses the order of the steps of photosynthesis. Because of the student's lack of prior knowledge, their mistake would likely not inform a strategic shift in how they encode the correct answer. The student would likely not be able to generate a coherent explanation for why their original ordering of the steps of photosynthesis was incorrect and why the correct ordering is biologically accurate (Callender & McDaniel, 2007). Thus, individuals with lower prior knowledge may benefit less, if at all, from generating errors relative to errorless learning, particularly when learning complex concepts.

Indeed, previous research has revealed that individuals with lower prior knowledge can benefit less learning from activities that introduce opportunities for mistakes (Carpenter et al., 2016; Cooper, Tindall-Ford, Chandler, & Sweller, 2001; Kalyuga, 2007; Leppink, Broers, Imbos, van der Vleuten, & Berger, 2012; Mulligan, Rawson, Peterson, & Wissman, 2018). For example, Carpenter and colleagues (2016) examined how practice tests affected student learning in a college biology class. Students learned about cell formation and then either retrieved key definitions and labeled diagrams from memory or copied definitions and diagram labels without engaging in any retrieval. Consistent with their prior performance in the course, students who were earning a higher grade in the course retrieved more correct definitions and diagram labels than students who were earning a lower grade in the course. On a subsequent quiz on the same material, higher-performing students benefited more from retrieval than copying. In contrast, lower-performing students benefited more from copying than retrieval. Thus, lower-performing students had lower prior knowledge, made more mistakes on the practice test, and were less likely to correct their mistakes on the subsequent quiz. Future research should experimentally manipulate prior knowledge and examine whether prior knowledge moderates the benefits of learning from errors with complex materials, such as biology definitions and diagrams.

## Conclusion

Although mistakes may be a frustrating part of learning something new, they should not be prevented; mistakes often create potent opportunities for new learning. A priority for memory research is to further understand how errors enhance learning in order to inform how learning experiences should be designed to maximize the likelihood that errors will be corrected across a broad range of students, content, and activities.

# REFERENCES

Akan, M., Stanley, S. E., & Benjamin, A. S. (2018). Testing enhances memory for context. *Journal of Memory and Language*, *103*, 19-27.

Anderson, J. R. (1996). *The architecture of cognition* (Vol. 5). Psychology Press.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic bulletin & review*, *7*(3), 522-530.

Anderson, N. D., & Craik, F. I. M. (2006). The mnemonic mechanisms of errorless learning. Neuropsychologia, 44, 2806–2813.

Ariel, R., & Dunlosky, J. (2013). When do learners shift from habitual to agenda-based processes when selecting items for study?. *Memory & Cognition*, *41*(3), 416-428.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: when agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432.

Ausubel DP. 1968. *Educational Psychology: A Cognitive View*. New York: Holt, Rinehart & Winston

Bandura A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval.

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin & Review*, *15*(1), 52-57.

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 529-533.

Bower, G. H., & Gilligan, S. G. (1979). Remembering information related to one's self. *Journal of Research in Personality*, *13*(4), 420-432.

Bridger, E. K., & Mecklinger, A. (2014). Errorful and errorless learning: The impact of cue–target constraint in learning from errors. *Memory & cognition*, *42*(6), 898-911.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review*, *18*(6), 1238-1244.

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

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology*, *99*(2), 339.

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

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*(2), 353-375.

Carpenter, S. K., & Vul, E. (2011). Delaying feedback by three seconds benefits retention of face–name pairs: the role of active anticipatory processing. *Memory & Cognition*, *39*(7), 1211-1221.

Chan, J. C., & McDermott, K. B. (2007). The testing effect in recognition memory: a dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431.

Cho, K. W., Neely, J. H., Brennan, M. K., Vitrano, D., & Crocco, S. (2017). Does testing increase spontaneous mediation in learning semantically related paired associates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(11), 1768–1778.

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

Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by imagining. Journal of Experimental Psychology: Applied, 7, 68–82.

Coppens, L. C., Verkoeijen, P. P., Bouwmeester, S., & Rikers, R. M. (2016). The testing effect for mediator final test cues and related final test cues in online and laboratory experiments. *BMC Psychology*, *4*(1), 25.

Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 599-607.

Cyr, A. A., & Anderson, N. D. (2012). Trial-and-error learning improves source memory among young and older adults. *Psychology and aging*, *27*(2), 429.

Cyr, A. A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 841.

Cyr, A. A., & Anderson, N. D. (2018). Learning from your mistakes: does it matter if you're out in left foot, I mean field?. *Memory*, *26*(9), 1281-1290.

de Jonge, M., Tabbers, H. K., Pecher, D., Jang, Y., & Zeelenberg, R. (2015). The efficacy of self-paced study in multitrial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 851.

DeLozier, S., & Rhodes, M. G. (2015). The impact of value-directed remembering on the own-race bias. *Acta Psychologica*, *154*, 62-68.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945-955.

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy production?. *Psychology and Aging*, *13*(4), 597.

Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition*, *29*(2), 247-253.

Dunlosky, J., Tauber, S., Mueller, M., & Thiede, K. (2016) Methodology for Investigating Human Memory: Problems and Pitfalls. In The Oxford Handbook of Metamemory: Oxford University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178-186.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88-92.

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 632.

Finn, B. (2017). A framework of episodic updating: An account of memory updating after retrieval. In *Psychology of learning and motivation* (Vol. 67, pp. 173-211). Academic Press.

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

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316.

Goldsmith, M., & Koriat, A. (2007). The strategic regulation of memory accuracy and informativeness. *Psychology of Learning and Motivation*, *48*, 1-60.

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

Halamish, V., Goldsmith, M., & Jacoby, L. L. (2012). Source-constrained recall: Front-end and back-end control of retrieval quality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 1.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement?. *Psychonomic Bulletin & Review*, *19*(1), 126-134.

Hausman, H., & Rhodes, M. G. (2018). When pretesting fails to enhance learning concepts from reading texts. *Journal of Experimental Psychology: Applied*, *24*(3), 331.

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.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269-299.

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

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology* (pp. 383-400). New York: The Guilford Press.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*(5), 852-857.

Jacoby, L. L., Shimizu, Y., Velanova, K., & Rhodes, M. G. (2005). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language*, *52*(4), 493-504.

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.

Jacoby, L. L., Wahlheim, C. N., & Kelley, C. M. (2015). Memory consequences of looking back to notice change: Retroactive and proactive facilitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1282.

Jacoby, L. L., Wahlheim, C. N., & Yonelinas, A. P. (2013). The role of detection and recollection of change in list discrimination. *Memory & Cognition*, *41*(5), 638-649.

Janes, J. L., Dunlosky, J., & Rawson, K. A. (2018). How do students use self-testing across multiple study sessions when preparing for a high-stakes exam?. *Journal of Applied Research in Memory and Cognition*, *7*(2), 230-240.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509-539.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469.

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?. *Memory*, *17*(4), 471-479.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-Based Learning: An Episodic Context Account. *Psychology of Learning and Motivation*, 237.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*(4), 704-721.

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731-746.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of experimental psychology: General*, *126*(4), 349.

Koriat, A. (2008). When confidence in a choice is independent of which choice is made. *Psychonomic Bulletin & Review*, *15*(5), 997-1001.

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*(5), 959-972.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490.

Koriat, A., Goldsmith, M., & Halamish, V. (2008). Controlled Processes in Voluntary Remembering. *Learning and Memory: A Comprehensive Reference*, 307–324.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36.

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 106.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219-224.

Kornell, N., & Finn, B. (2016). Self-regulated learning: An overview of theory and data. *The Oxford Handbook of Metamemory*, 325-340.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989.

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1573.

Leppink, J., Broers, N. J., Imbos, T., van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Self-explanation in the domain of statistics: an expertise reversal effect. Higher Education, 63, 771–785.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability in free and cued recall. *Bulletin of the Psychonomic Society*, *22*(6), 497-500.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective?. *Journal of Experimental Psychology: General*, *122*(1), 47.

McNamara, T. P. (2014). *Semantic Priming Perspectives from Memory and Word Recognition*. Abingdon, Oxon: Taylor and Francis.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465-489.

Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia*, *138*, 107296.

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*(4), 530.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463-477.

Metcalfe, J., & Miele, D. B. (2014). Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors. *Journal of Applied Research in Memory and Cognition*, *3*(3), 189-197.

Metcalfe, J., & Xu, J. (2018). Learning from one's own errors and those of others. *Psychonomic Bulletin & Review*, *25*(1), 402-408.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition.* New York, NY: Psychology Press. doi:10.4324/ 9780203338001

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257-271.

Moscovitch, M., & Craik, F. I. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior*, *15*(4), 447-458.

Mulligan, N. W., Rawson, K. A., Peterson, D. J., & Wissman, K. T. (2018). The replicability of the negative testing effect: Differences across participant populations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(5), 752.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, *10*(5-6), 389-395.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*(4), 207-213.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the" labor-in-vain effect.". *Journal of experimental psychology: Learning, Memory, and Cognition*, *14*(4), 676.

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The psychology of Learning and Motivation*, *26*, 125-141.

Peterson, D. J., & Wissman, K. (2020). Using tests to reduce mind-wandering during learning review. *Memory*, 1-6.

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*(6), 1051.

Putnam, A. L., Wahlheim, C. N., & Jacoby, L. L. (2014). Memory for flip-flopping: Detection and recollection of political contradictions. *Memory & Cognition*, *42*(7), 1198-1210.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335-335.

Raaijmakers, J. G. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, *27*(3), 431-452.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619-633.

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology*, *46*(2), 168-175.

Rhodes, M. G., & Kelley, C. M. (2005). Executive processes, memory accuracy, and memory monitoring: An aging and individual difference analysis. *Journal of Memory and Language*, *52*(4), 578-594.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432.

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review*, *21*(6), 1516-1523.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human learning and Memory*, *4*(6), 592.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99-115.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204.

Stevenson H, Stigler JW. 1994. *The Learning Gap: Why Our Schools Are Failing and What We Can Learn from Japanese and Chinese Education*. New York: Simon & Schuster

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 437.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109-118.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381-391.

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive Psychology*, *5*(3), 275-287.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory?. *Psychonomic Bulletin & Review*, *19*(5), 899-905.

Wahlheim, C. N. (2014). Proactive effects of memory in young and older adults: The role of change recollection. *Memory & Cognition*, *42*(6), 950-964.

Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory & Cognition*, *43*(1), 27-38.

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.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(4), 442.

Woloshyn, V. E., Pressley, M., & Schneider, W. (1992). Elaborative-interrogation and prior-knowledge effects on learning of facts. *Journal of Educational Psychology, 84*(1), 115–124.

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1073.

# APPENDIX A

Table A1

*Sample Materials for Experiments 1 and 2*

| Cue | Target |
|---|---|
| BACON | BREAKFAST |
| BATHROOM | TOWEL |
| CAREER | MONEY |
| CERAMIC | GLASS |
| GOWN | WEDDING |
| KITE | WIND |
| SAUCE | GRAVY |
| SWIM | FLOAT |
| POCKET | WALLET |
| ROPE | TWINE |
| MILITARY | UNIFORM |
| BIRD | FEATHER |
| CANDLE | WICK |
| CASE | LAWYER |
| DOLL | CHILD |
| JEWEL | STONE |
| SNAP | BUTTON |
| STREAM | CREEK |
| PEPPER | SPICE |
| OYSTER | PEARL |
| ASHES | SMOKE |
| DANDRUFF | SCALP |
| FRECKLE | MOLE |
| HYDROGEN | ATOM |
| WHALE | MAMMAL |
| RULER | INCH |
| BAIT | HOOK |
| GLUE | PAPER |
| KETTLE | WATER |
| POEM | SONG |

# APPENDIX B

The Kruskal-Goodman gamma correlation (Nelson, 1984) was used in Experiment 2 to compute confidence-accuracy resolution and accuracy-reporting resolution. Gamma is undefined for a participant when there is no variability in one or both of the dependent variables being correlated. Therefore, in Experiment 2, confidence-accuracy resolution was undefined for participants who either gave the same confidence rating for all selected answers or for whom the answers that they selected on step 2 were either all correct or all incorrect. Similarly, accuracy-reporting resolution was undefined for participants whose selected answers were either all correct or all incorrect or for participants who either reported all of their answers or withheld all of their answers on step 4. Table B1 displays the number of participants in each condition whose gamma correlations were undefined for each reason.

Table B1

*Number of Participants Excluded from Gamma Correlation Computations Due to Various Sources of Lack of Variance*

| Condition | Same Confidence Rating | All Selected Answers Correct | All Selected Answers Incorrect | Reported All Answers | Withheld All Answers |
|---|---|---|---|---|---|
| errorless | 2 | 2 | 0 | 5 | 0 |
| errorful+immediate | 0 | 1 | 0 | 2 | 0 |
| errorful+delayed | 0 | 0 | 0 | 2 | 0 |

# APPENDIX C

In Experiment 2, participants rated the likelihood that the answer they selected as the best candidate answer on step 2 of the final test was the correct answer. The primary purpose of collecting these confidence ratings was to calculate confidence-accuracy resolution, or the correlation between item-by-item confidence ratings and accuracy. Table C1 reports mean confidence ratings and calibration by condition, for completeness. For each participant, calibration was calculated as the difference between mean confidence rating for the answers selected on step 2 of the final test and the proportion of selected answers that were correct (i.e., forced-report accuracy). A positive calibration indicates that participants were overconfident, predicting that their selected answers were more likely to be correct than they actually were. A negative calibration indicates that participants were underconfident, predicting that their selected answers were less likely to be correct than they actually were (Rhodes, 2016).

Table C1

*Mean (SD) Confidence and Calibration Scores in Experiment 2*

| Condition | Confidence | Calibration |
|---|---|---|
| errorless | 63.06<br>(17.45) | 0.11<br>(0.14) |
| errorful+immediate | 76.77<br>(11.50) | 0.05<br>(0.14) |
| errorful+delayed | 69.11<br>(18.01) | 0.06<br>(0.30) |

Participants were somewhat overconfident with small differences in calibration among the three conditions, $F(2,166) = 3.14$, $p = .05$, $\eta_p^2 = .04$. Participants in the errorless condition were somewhat more overconfident than participants in the errorful conditions, although participants were well-calibrated overall.