Optimizing the Correction of Memory Errors

by

Hillary Gray Mullet

Department of Psychology & Neuroscience
Duke University

Date: ______________________

Approved:

___________________________

Elizabeth Marsh, Supervisor

___________________________

David Rubin

___________________________

Tobias Egner

___________________________

Gavan Fitzsimons

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2016
ABSTRACT

Optimizing the Correction of Memory Errors

by

Hillary Gray Mullet

Department of Psychology & Neuroscience
Duke University

Date: _________________________

Approved:

___________________________________________
Elizabeth Marsh, Supervisor

___________________________________________
David Rubin

___________________________________________
Tobias Egner

___________________________________________
Gavan Fitzsimons

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2016
Abstract

People are always at risk of making errors when they attempt to retrieve information from memory. An important question is how to create the optimal learning conditions so that, over time, the correct information is learned and the number of mistakes declines. Feedback is a powerful tool, both for reinforcing new learning and correcting memory errors. In 5 experiments, I sought to understand the best procedures for administering feedback during learning. First, I evaluated the popular recommendation that feedback is most effective when given immediately, and I showed that this recommendation does not always hold when correcting errors made with educational materials in the classroom. Second, I asked whether immediate feedback is more effective in a particular case—when correcting false memories, or strongly-held errors that may be difficult to notice even when the learner is confronted with the feedback message. Third, I examined whether varying levels of learner motivation might help to explain cross-experimental variability in feedback timing effects: Are unmotivated learners less likely to benefit from corrective feedback, especially when it is administered at a delay? Overall, the results revealed that there is no best “one-size-fits-all” recommendation for administering feedback; the optimal procedure depends on various characteristics of learners and their errors. As a package, the data are consistent with the spacing hypothesis of feedback timing, although this account does not successfully explain all of the data in the larger literature.
Contents

Abstract .............................................................................................................................. iv
List of Tables .................................................................................................................. vi
List of Figures ................................................................................................................ vii
Acknowledgements ...................................................................................................... viii
1. Introduction ................................................................................................................... 1
   1.1 But Errors Might Be Good For You ................................................................. 4
   1.2 Learning from Desirable Difficulties ............................................................. 7
   1.3 The Power of Feedback: A Brief History and Review .................................. 10
      1.3.1 Animal Learning, Behaviorism, and Early Classroom Studies ........... 10
      1.3.2 A Changing Zeitgeist and New Evidence on Feedback Timing .......... 14
      1.3.3 Feedback Timing: A Theoretical Perspective ....................................... 16
      1.3.4 Popular Views on Feedback Timing ...................................................... 20
   1.4 Overview of the Current Experiments .............................................................. 22
2. Experiments 1 and 2: Effects of Feedback Timing in Classroom Learning .......... 25
   2.1 Experiment 1: The Role of Feedback Timing in Classroom Learning: A Between-Subjects Manipulation with Student Impressions .................. 28
      2.1.1 Methods .................................................................................................... 28
         2.1.1.1 Participants ...................................................................................... 28
         2.1.1.2 Design ............................................................................................ 28
         2.1.1.3 Materials ....................................................................................... 28
         2.1.1.4 Procedure ...................................................................................... 29
2.1.2 Results

2.1.2.1 Homework Performance

2.1.2.2 Feedback Viewing

2.1.2.3 Exam Performance

2.1.2.4 Student Impressions

2.2 Experiment 2: The Role of Feedback Timing in Classroom Learning: A Within-Subjects Manipulation with Student Impressions

2.2.1 Methods

2.2.1.1 Participants

2.2.1.2 Design

2.2.1.3 Materials

2.2.1.4 Procedure

2.2.2 Results

2.2.2.1 Homework Performance

2.2.2.2 Feedback Viewing

2.2.2.3 Exam Performance

2.2.2.4 Exam Performance as a Function of Retention Interval

2.2.2.5 Student Impressions

2.3 Discussion

3. Experiments 3 and 4: Correcting Strongly-Held Errors: A Case for Immediate Feedback

3.1 Experiment 3
## 3.1 Methods

### 3.1.1 Methods

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>51</td>
</tr>
<tr>
<td>Design</td>
<td>51</td>
</tr>
<tr>
<td>Materials</td>
<td>51</td>
</tr>
<tr>
<td>Procedure</td>
<td>52</td>
</tr>
</tbody>
</table>

### 3.1.2 Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scoring and Analysis</td>
<td>54</td>
</tr>
<tr>
<td>Correct Answers</td>
<td>55</td>
</tr>
<tr>
<td>Inferences</td>
<td>55</td>
</tr>
</tbody>
</table>

### 3.2 Experiment 4

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>57</td>
</tr>
<tr>
<td>Participants</td>
<td>57</td>
</tr>
<tr>
<td>Design</td>
<td>58</td>
</tr>
<tr>
<td>Materials</td>
<td>58</td>
</tr>
<tr>
<td>Procedure</td>
<td>58</td>
</tr>
</tbody>
</table>

### 3.2.2 Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scoring and Analysis</td>
<td>59</td>
</tr>
<tr>
<td>Correct Answers</td>
<td>59</td>
</tr>
<tr>
<td>Inferences</td>
<td>60</td>
</tr>
</tbody>
</table>

### 3.3 Discussion

4. Experiment 5: Effects of Motivation on Feedback Processing and Learning From Feedback | 62 |
List of Tables

Table 1: Average proportion of correct multiple-choice responses on course exams in Experiments 1 and 2 as a function of short (1-2 weeks) versus long (3-4 weeks) retention interval ................................................................. 42

Table 2: Proportion of sentence fragments answered correctly versus with inferences or other wrong answers for the immediate feedback condition and two delayed feedback conditions of Experiment 3 ................................................................. 55

Table 3: Proportion of sentence fragments answered correctly versus with inferences or other wrong answers (Experiment 4) ................................................................. 62

Table 4: Proportion of Swahili-English word pairs correctly remembered on the initial test, presented separately for each of the six between-subjects conditions .................................................. 73

Table 5: Average time spent viewing the feedback message, presented separately for each of the six between-subjects conditions ................................................................. 75

Table 6: Proportion of Swahili-English word pairs correctly remembered on the final test .............................................................................................................................. 76

Table 7: Correlations between self-rated motivation and the key dependent variables in the experiment ........................................................................................................ 78

Table 8: Proportion of initially correct answers maintained on the final test .............. 80
List of Figures

Figure 1: The procedure for a classic spacing effect study (Panel A), and the typical procedure for a delayed feedback study (Panel B)................................................................................19

Figure 2: An example of an Immediate Feedback Assessment Technique (IF-AT) form............................................................................................................................................................................21

Figure 3: A sample problem from the experiment (left), pictured with the detailed solution that was provided as feedback (right)........................................................................................................29

Figure 4: Average number of times that a student viewed the feedback screen per homework problem..........................................................................................................................................................39

Figure 5: Proportion of correct multiple-choice responses on the course exams as a function of timing of feedback and feedback viewing condition.........................................................40

Figure 6: A schematic of the design for the immediate feedback condition and the two delayed feedback conditions (Experiments 3 & 4)................................................................................................53

Figure 7: A schematic of the procedure for Experiment 5............................................................72
Acknowledgements

There are many people who have helped and supported me throughout graduate school. First, I would like to thank my advisor, Elizabeth Marsh, who has always provided me with helpful advice and stimulating conversations about research ideas. Over the past five years, her kind and patient feedback has helped me to grow as a writer, communicator, and scientific thinker. My committee members, David Rubin, Tobias Egner, and Gavan Fitzsimons have also supported me with enthusiastic discussions and helpful brainstorming sessions. I would also like to thank Charles Brewer and Gil Einstein for sparking my interest in scientific research, introducing me to the fascinating study of human behavior, and helping me to develop the skills and confidence to succeed in graduate school. Finally, I am thankful for the steadfast support of my family and friends, who have always been there to encourage me through challenges and to celebrate my successes.
1. Introduction

Whether a foreign language learner practicing vocabulary, an instructor attempting to recall the names of new students, or a student contemplating the appropriate procedure for solving a math problem, people are always at risk of making errors when they attempt to retrieve recently learned information from memory. Such mistakes are avoidable if learners do not make active attempts to retrieve information during learning—if the foreign language learner, for example, consults a dictionary rather than independently attempting to recall the meaning of a word. An important and interesting question is which practice is more beneficial for learning. That is, should learners be encouraged to actively produce answers from memory, thereby guaranteeing that they will sometimes make errors? Or, is it better to learn by rereading, studying, or reviewing the correct information, protecting oneself from making mistakes?

The question of how generating errors affects learning has long been a focus of empirical study, and there are many findings that caution against introducing errors into the learning process. This perspective likely originates in the animal learning literature of the early to mid-1900s. For example, Terrance (1963) taught pigeons to produce a key-pressing response to a specific stimulus through either standard discrimination training or an approach called *early progressive training*, which allowed the animals to learn the

---

desired response with few to no errors. Early progressive training led to better retention of the desired behavior, leading Terrance (1963) to conclude that the best learning conditions involve the avoidance of errors. Guthrie (1952) also endorsed this idea with his claim that animals “learn what they do.” From this perspective, whenever an animal makes an error, that incorrect stimulus-response association is stamped in, potentially hindering that animal’s ability to later learn and remember the correct response.

Further research with clinical human populations also supported the idea that making errors during learning can be harmful. For example, Baddeley and Wilson (1994; Wilson, Baddeley, Evans, & Shiel, 1994) taught memory-impaired participants² a series of words according to either an errorless or errorful learning procedure. On errorless learning trials, participants were simply told the correct word (e.g., “I am thinking of a six-letter word that starts with ‘AR,’ and the word is ‘ARTIST’”). On errorful trials, however, participants were asked to guess several possible completions of the word (e.g., “I am thinking of a six-letter word that starts with ‘AR’. Can you guess what it might be?”) before being told the correct answer, thereby ensuring that they would make errors. Critically, errorless learning benefited performance on a stem completion test given at the end of the experiment session (see also Hunkin, Squires, Parkin, & Tidy, 1998; Page, Wilson, Shiel, Carter, & Norris, 2006). To explain this result, the authors argued that

² The 16 patients were all classified as severely memory-impaired and had suffered from a variety of injuries and conditions, including Herpes Simplex Encephalitis, Korsakoff’s syndrome, and severe closed-head injury.
errors—once they have been produced—are maintained in implicit memory and are likely to come to mind again later. In the absence of a functioning explicit memory system (as experienced by amnesics), learners are unable to overcome the competition produced by those errors. For this reason, Baddeley and Wilson (1994) suggested that it is best to prevent errors from forming in the first place.

Other research has supported the claim that errors are best avoided even in healthy young adults with intact explicit memory. Making an error on an initial test strongly predicts the likelihood of making an error on that same item later on (e.g., Butler & Peterson, 1965); and learners often repeat the same errors over and over (Marx & Witter, 1972), even with intervening opportunities to restudy the correct information (Kay, 1955). Errors that are initially committed with high confidence are likely to reassert themselves one week later, even when learners have received corrective feedback (Butler, Fazio, & Marsh, 2011). One interpretation of these findings appeals to interference theory, which states that new and old information interact in memory (e.g., Underwood, 1945). For example, in the AB-AD interference paradigm, participants first learn a particular stimulus-response association (e.g., turtle—grape) and are later asked to update to a new association (e.g., turtle—magician) that shares the same stimulus term. Critically, on a final test on which learners are given the stimulus term and are asked to recall the most-recently learned response (magician), the older response (grape) sometimes intrudes instead. The AB-AD interference paradigm can be extended to
explain the detrimental effects of producing an error on later memory. Specifically, an error is like the B term in an AB-AD association; when the learner attempts to retrieve the correct answer, the error comes to mind instead.

1.1 But Errors Might Be Good for You: Effortful Retrieval and Desirable Difficulties

Although a substantial body of research has suggested advantages of avoiding errors during learning, this conclusion conflicts with other well established principles in the memory literature. For example, many errorless learning techniques seem to promote relatively passive learning, despite a long history of evidence that active and effortful encoding benefits memory (Craik, 1983, 1986; Cyr & Anderson, 2012). Indeed, the idea of introducing desirable difficulties into the learning process has received a recent emphasis in the memory and education literature (Bjork & Bjork, 2011; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Goode, Geraci, & Roediger, 2008; Roediger & Karpicke, 2006). Proposed by Bjork (1994a, 1994b; Christina & Bjork, 1991; Schmidt & Bjork, 1992), the desirable difficulties framework challenges the commonly held assumption that conditions that make learning easier and more efficient (as is often the case in errorless learning procedures) also produce the best long-term retention and transfer. Instead, drawing on Estes’ (1955) distinction between the momentary strength of a response and its long-lasting habit strength, the framework states that the conditions that increase the temporary accessibility of information and skills in memory (i.e., retrieval strength) are different from those that promote an enduring memory representation (i.e.,
storage strength). In other words, strategies that rapidly increase the retrievability of a response often actually lead to poorer long-term retention and transfer, in comparison to other strategies that make the learning process slower and more challenging.

Consistent with the desirable difficulties predictions, a large body of research has shown that conditions that make the learning process more effortful often produce better long-term memory. For example, the spacing effect—a benefit of practice that is distributed in time, rather than massed all at once—was first demonstrated by Ebbinghaus (1885) and has since been replicated hundreds of times (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a meta-analysis of the verbal learning studies). Spaced practice is often perceived by learners to be less effective (Baddeley & Longman, 1978; Dunlosky & Nelson, 1994; Kornell, 2009; Kornell & Bjork, 2008; Zechmeister & Shaughnessy, 1980) than massed practice, and it results in slower learning of the correct information (e.g., Roediger & Pyc, 2012). Nonetheless, these more difficult learning conditions typically produce better memory for learned information over time.

Support for the desirable difficulties framework also comes from the finding that the effortful strategy of taking a test functions as a powerful learning event (McDaniel, Anderson, Derbish, & Morisette, 2007; Roediger & Karpicke, 2006a, 2006b). Learners often study by reviewing or rereading material, rather than by attempting to retrieve that information from memory themselves (Karpicke, Butler, & Roediger, 2009). The effortful process of retrieval, however, has been shown to produce much larger benefits
for long-term retention and transfer (Roediger & Karpicke, 2006a, 2006b). Again, although retrieval can be difficult and is often perceived by learners to be less effective than other study strategies, these challenging learning conditions ultimately promote a more durable and flexible memory representation (Karpicke & Blunt, 2011; Karpicke & Roediger, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006a, 2006b).

Indeed, recent research has revealed that challenging learning conditions can benefit later memory even when many errors are produced during learning. For example—as long as feedback is provided—introducing spaced practice into learning has been shown to enhance long-term memory of foreign language word pairs and obscure English vocabulary even when the delay between tests is so long that errors occur on the majority of trials (Pashler, Zarow, & Triplett, 2003). In fact, Bahrick and Hall (2005) suggested that errors might actually be instrumental in producing the benefits of spaced practice. According to their metacognitive explanation of the spacing effect, the long retention-intervals under spaced practice guarantee that learners will sometimes fail, and this failure allows them to learn which strategies are effective for memory and which are not. In other words, retrieval failures (i.e., errors) can be important for learning.

Similarly, the active process of retrieving or generating information—as compared to simply studying—has been shown to benefit later memory, even when the generation attempt is rarely successful. For example, Kane and Anderson (1978) asked participants to guess the final word in sentences like “The dove appeared when the
magician said ____.” Although participants only guessed the correct answer (peace) on 9% of trials, they still benefited from the generation attempt. Slamecka and Fevreiski (1983) similarly created conditions in which generation was highly likely to fail; participants were asked to complete difficult antonym pairs like trivial-v____ (correct answer: vital). Generation was only successful 20% of the time, but still produced better retention of the correct answers as compared to studying. More recent research has shown that benefits of retrieval (i.e., testing effects) emerge in sixth grade and college classrooms even when a substantial number of errors are produced (Metcalf & Kornell, 2007). The take-home message is that testing, followed by corrective feedback, enhances learning—even when learners produce errors on many trials (see also Grimaldi & Karpicke, 2012; Huels & Metcalfe, 2012; Kang, Pashler, Cepeda, Rohrer, & Carpenter, 2011; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, 2014; Kornell, Hays, & Bjork, 2009).

1.2 Learning from Desirable Difficulties: Feedback is Necessary

As reviewed above, a large body of research shows that introducing challenges into the learning process often benefits learning, even when learners make many errors. A critical caveat, however, is that learners must receive corrective feedback during the learning process in order to improve their performance. Without feedback, learners are highly unlikely to correct their errors, as shown in experiments that included a challenging learning task and both (1) a group of participants that received corrective
feedback and (2) a no-feedback control condition (e.g., Fazio, Huelser, Johnson, & Marsh, 2010; Pashler, Cepeda, Wixler, & Rohrer, 2005). For example, Fazio et al.’s (2010) participants studied prose passages about topics like Alaska and then took an initial and a final test. Critically, some participants received feedback (i.e., another presentation of the correct answer) after providing their responses on the initial test trials, whereas others did not. Learners who received feedback gained an average of 30 percentage points on the final test, indicating that the feedback messages helped them to correct their errors, learn new correct information, or both. On the other hand, learners who did not receive feedback lost an average of 6 percentage points across tests. Pashler et al. (2005) reported a similar pattern: whereas learners corrected almost none of their incorrect foreign-language translations in a no-feedback control condition, the provision of feedback improved final test performance by 494%. In sum, feedback is critical for correcting errors, which occur often when learning conditions are challenging. Indeed, with an average effect size in the medium to large range ($d = 0.73$), feedback ranks among the most efficient and powerful tools for improving student learning (Hattie, 2009).

Although the power of feedback is well recognized, its exact definition— and the precise circumstances under which it should be administered— are not yet fully understood. The lack of clarity surrounding the concept of feedback can be seen in the following remark by Hattie (2009, p. 173):
When I completed the first synthesis of 134 meta-analyses of all possible influences on achievement it soon became clear that feedback was among the most powerful influences on achievement. Most programs and methods that worked best were based on heavy dollops of feedback. When I was presenting these early results in Hong Kong, a questioner asked what was meant by feedback… I have struggled to understand the concept of feedback ever since.

In other words, feedback is regarded as an important learning tool, but researchers and educators do not have a complete grasp on what it is or when to use it. Upon reflection, Hattie (2009) proposed the following definition: “feedback is information provided by an agent (e.g., a teacher, peer, book, parent, or one’s own experience) about aspects of one’s performance or understanding” (p. 174). Other definitions in the literature generally agree with this one; for example, Shute (2008) stated that feedback is “information communicated to the learner that is intended to modify his or her thinking for the purpose of improving learning” (p. 154).

As far as how and when to administer feedback, answers to some questions can be found in the existing literature. It is clear, for example, that an effective feedback message must do more than communicate to learners whether they are right or wrong; instead, it must also tell them the correct response that they should have provided instead. Indeed, correct/incorrect feedback does almost nothing to help learners correct their errors, and sometimes is not any better than no feedback at all (e.g., Fazio et al., 2010; Pashler et al., 2005). The only exception to this rule occurs when learners’ errors were made on a two-alternative forced choice test (e.g., True/False). In this case, correct/incorrect feedback is just as effective as correct answer feedback, because
learners are easily able to infer the correct answer (Marsh, Lozito, Umanath, Bjork, & Bjork, 2012).

Many other questions about how to administer optimal feedback procedures remain to be answered, however. Perhaps one of the most interesting and oft-debated of these issues concerns feedback timing. The question is whether feedback should always be administered immediately after a learner’s response, or if it is just as effective when it is given after a delay. A common belief shared by educators, students, and researchers is that feedback should always be administered as soon as possible. In the next section of this dissertation, I discuss the origins of this belief as well as some evidence for and against it. This discussion is couched within a broader overview of the feedback literature, including how the study of feedback began and has evolved over time. The goal is to set the stage for the experiments of my dissertation, which focus on understanding how to administer feedback to ensure the optimal conditions for correcting memory errors, with a specific focus on feedback timing.

1.3 The Power of Feedback: A Brief History and Review of the Literature

1.3.1 Animal Learning, Behaviorism, and Early Classroom Studies

The earliest systematic studies of feedback were animal learning experiments. Edward Thorndike, in a series of classic studies, trained hungry cats to escape from “puzzle boxes” to obtain food rewards. Observing the behavior of these animals led Thorndike to propose several theoretical laws of behavior, including the now-famous
Law of Effect, which states that, over time, organisms learn to repeat behaviors that previously led to a desired state of affairs, while they become less likely to repeat behaviors that previously produced unsatisfying outcomes (Thorndike, 1898, 1911). Although Thorndike would later revise the second part of this law (deemphasizing the importance of unsatisfying outcomes, or punishment), his ideas played a key role in shaping the notion that reinforcement and motivation are important components in the learning process. As a result of his work, researchers began to see feedback as a motivator of behavior.

With the rise of behaviorism in 1920s, researchers became eager to strip psychology of introspection and any explanations of behavior that invoked unobservable mental processes. Behaviorists rejected any speculation of what constituted “a satisfying state of affairs”; however, they retained Thorndike’s basic view of feedback as a reinforcer of behavior (e.g., Hull, 1943; Skinner, 1938; Thorndike, 1932). This perspective guided Skinner’s proposal that learning occurs when organisms begin to associate certain stimuli with particular consequences. Skinner demonstrated this principle using operant conditioning, a procedure through which animals are gradually guided towards a desired response by the reinforcement of behaviors that more and more closely approximate the ideal. Perhaps the most famous example involves Skinner’s success at training rats to press a bar inside a small glass box (an operant conditioning chamber or Skinner box) in order to receive food. To achieve this goal, Skinner first
reinforced the rats’ behavior with the delivery of food pellets whenever they turned
towards the bar; then he reinforced them only for touching the bar; and then finally only
for producing the desired barpressing behavior. By reinforcing this series of small,
discrete steps, the barpressing response could be shaped within minutes. An important
caveat, however, was that the food (i.e., feedback) had to be provided immediately after
each response; otherwise, the animal failed to perceive the contingency for learning (for a
review, see Renner, 1964). In fact, even a slight delay of feedback dramatically decreased
its effectiveness; in Skinner (1954)’s own words, “…the lapse of only a few seconds
between response and reinforcement destroys most of the effect” (p. 91, see also
Saltzman, 1951).

As psychologists mastered the art of using reinforcement to shape animal
behavior, they began to turn their gaze to the classroom with the question of whether the
same principles could be applied to enhance student learning. Behaviorists viewed the
goal of education as one of shaping desired responses; children were organisms in need
of reinforcement and parents and teachers were modifiers of behavior. Viewing the
education system through this lens led to two troubling observations. First, although
animal-learning studies demonstrated that optimal learning occurred when organisms
received step-by-step guidance through the learning process, students rarely received
individualized instruction or programming. Second, despite clear evidence that animals
learned best when feedback was given immediately, feedback in the classroom was often
provided after a lengthy delay (if it was given at all). Regarding the latter point, Skinner (1954, p. 191) made the following observation after a visit to his daughter’s fourth-grade classroom:

In many cases—for example, when papers are taken home to be corrected—as much as 24 hours may intervene [before students receive feedback]. It is surprising that this system has any effect whatsoever.

In response to these perceived problems, behaviorists and educational psychology researchers endeavored to develop new technologies and teaching approaches that incorporated important ingredients from animal learning studies: the use of small, discrete steps and the administration of immediate feedback. For example, both Sidney Pressey and Skinner created teaching machines, which worked by presenting learners with a question, requiring them to write a response, and then allowing them to reveal the correct answer by sliding a panel on the side. Importantly, these machines allowed learners to move through a lesson or set of problems at their own pace, being reinforced with corrective feedback at every step along the way. This technology was widely accepted during the 1960s, with more than 100,000 of Skinner’s machines being sold door-to-door (Benjamin, 1988). Another popular educational initiative was the token economy system, in which children received small “tokens” like stickers or fake money as a reward for good performance. Later, the children could exchange the tokens for a larger prize (e.g., a special snack or a small toy). The key point was that the tokens provided

---

3 Pressey’s machines were less successful, perhaps because they appeared during the Great Depression.
immediate reinforcement for the desired response; children did not have to wait to receive feedback when they performed well.⁴

1.3.2 A Changing Zeitgeist and New Evidence on Feedback Timing

In the 1950s and 60s, psychology experienced the rise of a new zeitgeist, and the conceptualization of feedback began to change. Psychologists became dissatisfied with behaviorism’s strict focus on stimulus-response relationships, which prohibited the examination of interesting questions about how organisms attend, perceive, think, and remember. During the cognitive revolution, behaviorism was discarded for a broader approach that also acknowledged internal characteristics of the mind. As part of this shift, psychologists began to view feedback as not just a mechanism for reinforcement, but also a tool for learning.

At the same time, a series of studies began to call into question the assumption that delaying feedback is always detrimental for learning (e.g., Anderson, Kulhavy, & Andre, 1972; Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962; Butler & Roediger, 2008; Carpenter & Vul, 2011; Metcalfe, Kornell, & Finn, 2009; Kulhavy & Anderson, 1972; More, 1969; Phye & Andre, 1981; Rankin & Trepper, 1978; Sassenrath & Yonge, 1968, 1969; Smith & Kimball, 2010; Surber & Anderson, 1975; Sturges & Crawford, 1964). For example, Brackbill and Kappy (1962) asked third graders to guess which of two line drawings was “correct” (as

⁴ Token economies are still used today, especially as part of behavior modification programs for children with various disorders including ADHD and autism.
arbitrarily designated by the experimenter); the correct answers were revealed 0, 5, or 10 seconds after learners made their responses. On a final test 24 hours later, learners who had received delayed feedback remembered more drawings than those who had received immediate feedback. In a more recent study, Butler, Karpicke, and Roediger (2007, Experiment 2) had participants read a set of expository passages and then take a test with feedback provided either immediately after each response or after a delay of 24 hours. On a final retention test one week later, participants performed better on items for which they had received delayed feedback.

How could such data be reconciled with earlier animal learning studies that had provided evidence in favor of immediate feedback? To explain the discrepant results, researchers pointed to differences in the mental abilities of their subjects—for example, the fact that humans but not animals have an ability to rehearse the cue (e.g., Brackbill & Kappy, 1962) and anticipate the correct response over the delay (e.g., Carpenter & Vul, 2011). Sturges (1978) speculated that “after a longer delay interval, students engage in a more thorough semantic analysis of the information presented at feedback…” (p. 386). Similarly, Atkinson (1968) proposed that inserting the delay between a response and subsequent feedback creates an opportunity for “abstracting out critical features that may facilitate associating the stimulus with the correct response when it is subsequently presented” (p. 8). In any case, the emergence of these new findings made it clear that, unlike animals, humans do not always need immediate reinforcement to learn new
information. Indeed, rather than stunting the learning process, delaying feedback often enhanced later memory.

1.3.3 Feedback Timing: A Theoretical Perspective

As the literature in support of delaying feedback continued to expand, two theories emerged that attempted to account for these findings. According to the interference perseverance hypothesis, delaying feedback is helpful because it allows learners to forget their errors (Kulhavy & Anderson, 1972). More specifically, the theory assumes that errors are a source of proactive interference, similar to a B term in a classic AB-AC (e.g., knee-bone—knee-bend) interference paradigm. According to the theory, learners’ errors exert less interference as they are forgotten over time; thus, the correct response is more easily learned and remembered when the feedback message is presented after a delay.

Although the interference preservation hypothesis provides a plausible and parsimonious explanation for the benefits of delaying feedback, there are several points of evidence that argue against it. For example, re-presenting learners with their initial errors as part of the delayed feedback message does not impair later memory for the correct information (e.g., Butler et al., 2007; Butler & Roediger, 2008). In addition, learners are often quite good at remembering their errors, regardless of whether feedback is provided immediately or after a delay (e.g., Peeck & Tillema, 1978, 1979; Peeck et al., 1985; Vaughn & Rawson, 2012). In fact, remembering an initial error has even been shown to correlate with better memory for the correct information (Butler, Fazio, &
Marsh, 2011), and there is some evidence that errors can be used as mediators, or pieces of a “memory chain” that help to remind learners of the correct responses (Huelser & Metcalfe, 2012; Pyc & Rawson, 2010, 2012). Finally, the interference preservation hypothesis predicts that—regardless of feedback timing—it would be better not to make an error at all than to make an error and receive corrective feedback; there is no interference to overcome if an error never occurs in the first place. Many studies show, however, that making an initial error often benefits later memory relative simply studying the to-be-learned information (e.g., Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012). Each of these pieces of evidence argues against the view remembering one’s errors is harmful—and, in turn, undermines the hypothesis that delaying feedback is beneficial because it allows learners to forget their mistakes.

Recently, three independent teams of researchers proposed a different account, the spacing hypothesis, to explain the beneficial effects of delaying feedback (Butler, Karpicke, & Roediger; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Smith, Kimball, & Mann, 2007). This account appeals to the well-known finding that distributing study episodes over time (i.e., spacing) is beneficial for long-term memory (for a review see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). According to the spacing hypothesis, delaying feedback provides an opportunity for spaced review of any items that were initially answered correctly; in contrast, immediate feedback provides massed exposure
to those same items. Thus, the spacing hypothesis states that delaying feedback improves memory by increasing the probability of maintaining one’s initially correct responses.

Note that this statement contains an interesting assumption of the spacing hypothesis: delaying feedback should only benefit memory when learners produce at least some correct information on the initial test.

Smith and Kimball (2010) reasoned that, if the benefits of delaying feedback are in fact due to a kind of spacing effect, then certain patterns that occur in the broader literature on spacing should also appear when examining data regarding feedback timing (Cepeda et al., 2006; Dempster, 1996; Hintzman, 1974). One such pattern involves a well-documented relationship between the length of time that separates repeated study sessions (i.e., the interstudy gap) and the length of time that the studied information is expected be remembered (i.e., the retention interval; see Panel A of Figure 1). More specifically, analyses of hundreds of studies have revealed that when learners are expected to remember information for a short amount of time, the optimal gap between study sessions is shorter than when they are expected to remember that information for a longer time period (Balota, Duchek, & Paullin, 1989; Cepeda et al., 2006, 2009; Glenberg, 1976; Peterson, Wampler, Kirkpatrick, & Saltzman). The general “rule” is that final-test memory is optimized when the interstudy gap is about 10%-20% of the retention interval. For example, learners who want to remember a set of information for a total of 1-week should space their study sessions approximately 1 day apart; when the
critical information must be remembered for six months, however, a 28-day inter-study gap is optimal (Pashler, Rohrer, Cepeda, & Carpenter, 2007).

Figure 1: Panel A shows the procedure for a classic spacing effect study, in which two or more exposures to the critical information are separated by a certain time interval (i.e., the interstudy gap). The retention interval is the time period separating the last spaced exposure from the final test. Panel B shows that delayed feedback studies also contain an interstudy gap and retention interval.

Smith and Kimball (2010) asked whether a similar relationship between interstudy gap and retention interval occurs in studies involving delayed feedback (Panel B of Figure 1). In other words, does the optimal gap between an initial test and the administration of feedback depend on how long learners are expected to remember the critical information? To test this question, they created a series of schedules for learning 96 trivia facts. For each item, participants completed an initial study session, an initial
test, a feedback review phase, and final test. The initial study session always took place on Day 1 of the experiment, and the final test was on Day 8. The critical manipulation was the length of time between the initial test and the feedback review phase, which varied across items and was 8 minutes, 24 hours, or 48 hours. Importantly, and consistent with the rule from the broader spacing effect literature that the ideal interstudy gap is about 10-20% of the final retention interval, the 24 hour interstudy gap (which represented 13% of the 8-day retention interval) produced the best final test performance. This result provides evidence consistent with the spacing hypothesis as an explanation for the benefits of delaying feedback on later memory.

1.3.4 Popular Views on Feedback Timing

Despite a substantial (and growing) body of evidence for potential benefits of delaying feedback, the majority of students and educators continue to endorse the superiority of immediate feedback. This view can be seen in various course management and classroom assessment tools that advertise the delivery of immediate feedback as one of their primary selling points. For example, the Immediate Feedback Assessment Technique (IF-AT; Epstein Educational Enterprises, 2013; see Figure 2) is a popular tool that allows students to obtain corrective feedback in real-time as they complete a multiple-choice exam. Each multiple-choice alternative is covered by a thin opaque film, similar to the covering on a lottery ticket. Students must “scratch off” their first-choice response in order to reveal a star or other symbol that indicates correctness. If their first
answer is not correct, they can continue scratching off additional response options with the possibility of earning partial credit. According to the IF-AT website, this tool “is based on solid psychological principles, [including that] immediate feedback is beneficial for learning (and is superior to delayed feedback).”

Figure 2: An example of an Immediate Feedback Assessment Technique (IF-AT) form.

This strong endorsement of immediate feedback is unsurprising when considered in the context of a larger pattern, in which people often believe that the most effective learning conditions are those that feel easiest and most fluent. For example, retrieval practice—the act of bringing a piece of learned information to mind, without consulting an outside source like a textbook or notes—is a robust and highly effective strategy for
improving learning and memory (Roediger & Karpicke, 2006a, 2006b). In part because this strategy feels difficult, however, learners tend to discount its effectiveness, instead favoring less effective strategies like highlighting, reading, and note-taking (which tend to produce a greater feeling of fluency; Karpicke, 2009; Karpicke, Butler, & Roediger, 2009). Learners are also tend discount the benefits of spacing, which produces feelings of disfluency and difficulty, instead rating massed practice as being more effective (Kornell, 2009). In the broader context of these types of metacognitive illusions, it is unsurprising that learners and educators continue to advocate for immediate feedback, which feels easier and more fluent than receiving feedback after a delay.

1.4 Overview of the Current Experiments

The goal of my dissertation was to shed light on several open questions regarding the optimal procedures for administering feedback, with a specific focus on feedback timing. The first two experiments were conducted in real college classrooms in which students learned complex engineering concepts. The primary goal of these experiments was to examine whether benefits of delaying feedback could be replicated in “noisy” classroom environments; as described in more detail in the following section, there is some speculation from reviewers of the feedback literature about whether benefits of delaying feedback—which have been replicated repeatedly in laboratory contexts—generalize to applied settings like the classroom. A second interest was in documenting the metacognitive beliefs of students who received immediate or delayed feedback on
homework assignments throughout the semester. Would students have an accurate sense of how the feedback-timing procedures affected their learning?

Experiments 3 and 4 took a different focus, examining the correction of a specific type of error for which it seemed likely that delaying feedback would not be beneficial. Although debates regarding the optimal feedback-timing procedure have been plentiful, there has been little to no discussion about whether the best time to administer feedback depends on the type of error to be corrected. Experiments 3 and 4 examined the formation and subsequent correction of false memories, which are strongly held errors that tend to share a semantic relationship with the correct information (Roediger & McDermott, 1995; Sampaio & Brewer, 2009). These types of errors may often go undetected by the learner, thus partially explaining why they are difficult to correct. Experiments 3 and 4 examined the prediction that false memories would be corrected at higher rates when the learning conditions specifically drew attention to the contrast between learners’ errors and the correct information, as is the case with immediate feedback that can be compared back-to-back with one’s error. In terms of the larger picture of optimizing error correction, Experiments 3 and 4 highlight the fact that the same recommendations for administering feedback do not necessarily apply across all circumstances; in particular, strongly held errors may go unnoticed by the learner unless the correction conditions specifically highlight them.
Finally, Experiment 5 examined the relationship between feedback timing and another variable that is assumed to play a key role in learning from feedback: learner motivation (Butler & Winne, 1995; Hattie & Timpersley, 2007; Mory, 2004; Shute, 2008). A basic assumption is that learners must be reasonably motivated in order to attend to and process a feedback message; without motivation, feedback will have no effect on later memory (e.g., Hancock, Thurman, & Hubbard, 1995). Extrapolating from this hypothesis, one can make a further prediction regarding feedback timing. Specifically, modulations in learner motivation may have a particularly pronounced effect on learning when feedback is delayed, because the unmotivated learner may be especially unlikely to process a delayed feedback message. The goal of Experiment 5 was to test this hypothesis. As the whole, the major objective of my dissertation was to gain a more nuanced understanding of how to administer feedback in a way that benefits memory. Returning to the point from the beginning of this chapter, making errors is rarely harmful (and, in fact, can even be beneficial), but is important to understand how best to correct those mistakes.
2. Experiments 1 and 2: Effects of Feedback Timing in Classroom Learning

As discussed above, many laboratory studies have provided support for the notion that delaying feedback is beneficial for learning. Despite acknowledging these data, however, reviewers of the feedback literature have expressed skepticism that such results generalize outside of the tightly controlled conditions of the laboratory. For example, based on a meta-analysis of studies that manipulated feedback timing, Kulik and Kulik (1988) concluded that delayed feedback is often superior to immediate feedback in the laboratory, but immediate feedback is better in the classroom. Mory reached a similar conclusion, stating that “…in most [real-life] learning situations delayed feedback appears to function to hinder the acquisition of needed information” (p. 930).

The authority of these claims is dampened, however, upon taking a closer look at the studies on which these reviews are based. Specifically, many of the included studies that show an advantage of immediate over delayed feedback contain one or more serious methodological flaws that undermine their conclusions (Angell, 1949; Little, 1934; Pressey, 1950; Sullivan et al., 1971). For example, some studies confounded the type and timing of feedback (Little, 1934; Sullivan et al., 1971), the number of times each group received feedback (Paige, 1966), or included both of these confounds (Angell, 1949).

---

Other studies measured speed of acquisition (i.e., how many trials participants took to learn the correct responses) rather than long-term retention (Landsman & Turkewitz, 1962; Markowitz & Renner, 1966; Saltzman, 1951; Sturges et al., 1968). Because factors that increase the rate of initial learning often fail to improve long-term memory, these studies cannot speak to the question of how feedback timing affects retention (Bjork & Bjork, 1992). There is, however, one methodologically sound study that shows an advantage of immediate feedback in the classroom. White (1968) assigned a series of eight-short tests to students in a semester-long educational psychology course. Students reviewed the correct answers either the same day as each exam (immediate feedback) or two days later (delayed feedback). On a final test containing different problems at the end of the semester, students who had received immediate feedback performed better.

Given lack of consensus in the existing literature, there is a need to further examine the effects of feedback timing in the classroom. Experiments 1 and 2 were conducted during successive semesters of an upper-level undergraduate engineering course (“Continuous Time Signals and Systems”) at University of Texas at El Paso (UTEP). The instructor gave two lectures each week and assigned students additional materials for viewing outside the classroom (e.g., texts and videos). Students completed the weekly homework assignments outside of class through OpenStax Tutor, an e-learning system designed to allow experimentation and the implementation of cognitive science principles to benefit student learning (Butler, Marsh, Slavinsky, & Baraniuk,
Feedback on the homework problems was released to students either immediately after the assignment deadline or one week after the deadline. Importantly, the feedback message was identical for all students regardless of the timing of feedback. Students were required to view the feedback in order to receive credit for the assignment in Experiment 1, and this requirement was manipulated in Experiment 2.

Student learning from the feedback was assessed on course exams, which required students to transfer their knowledge of the concepts covered in the homework assignments to novel problems. Students also completed a survey about their perceptions of the relative effectiveness of immediate and delayed feedback as well as their personal preferences on the optimal timing of feedback. Importantly, the surveys were conducted at the end of the semester so that students’ responses could be informed by their experiences throughout the course. In addition, the timing of feedback was manipulated between-students in Experiment 1 and within-students in Experiment 2, which allowed a comparison between the perceptions of students who experienced only immediate or delayed feedback in the course with those of students who experienced both timing conditions.
2.1 Experiment 1: The Role of Feedback Timing in Classroom Learning: A Between-Subjects Manipulation with Student Impressions

In Experiment 1, timing of feedback was manipulated across two sections of the course. Students in one section received feedback immediately after the assignment deadline, whereas students in the other section received it one week after the deadline.

2.1.1 Methods

2.1.1.1 Participants

Twenty-six students consented to release their data for research purposes.

2.1.1.2 Design

Timing of feedback (immediate vs. delayed) was manipulated between-subjects.

2.1.1.3 Materials

A total of 11 homework assignments were assigned using OpenStax Tutor. Three of these assignments were excluded from the final data analysis because they took place during the week immediately preceding an exam, and—in order to ensure fairness—all students received immediate feedback. Each homework assignment contained 10-14 practice problems corresponding to 3-6 core concepts that were covered in lecture during the week preceding the assignment. Each practice problem required the application of a concept to determine the solution (see Figure 3 for an example). A different set of problems was used for the course exams. Like the homework problems, each exam problem required the application of a concept. Each exam contained 5-10 problems with
approximately equal coverage of the concepts from the preceding homework assignments (due to time limitations, the exams did not contain a problem for every concept).

Find the unilateral Laplace transform $X(s)$ of signal $x(t) = \sin(2\pi t)u(t)$, where $u(t)$ is the unit step function.

- **a)** $X(s) = \frac{2\pi}{s^2 + 4\pi^2}$
- **b)** $X(s) = \frac{2\pi}{s^2 - j2\pi}$
- **c)** $X(s) = \frac{2\pi}{s^2 - 2\pi}$
- **d)** $X(s) = \frac{2\pi}{s^2 - 4\pi^2}$
- **e)** $X(s) = \frac{2\pi}{s^2 - 4\pi^2}$

**Detailed Solution**

Using the Laplace transform integral, we have

$$X(s) = \int_0^\infty x(t) e^{-st} dt$$

$$= \int_0^\infty \sin(2\pi t) e^{-st} dt$$

From the trigonometric identity $\sin(2\pi t) = \frac{e^{j2\pi t} - e^{-j2\pi t}}{2j}$, we can write

$$= \int_0^\infty \frac{e^{j2\pi t} - e^{-j2\pi t}}{2j} e^{-st} dt$$

$$= \int_0^\infty \frac{e^{j2\pi t}}{2j} e^{-st} dt - \int_0^\infty \frac{e^{-j2\pi t}}{2j} e^{-st} dt$$

$$= \frac{1}{2j} \left[ \frac{e^{j(\pi - 2\pi)}}{s - j2\pi} \right]_0^\infty - \frac{1}{2j} \left[ \frac{e^{-j(\pi + 2\pi)}}{s + j2\pi} \right]_0^\infty$$

Assuming $\text{Re}[s] > 0$, then we have

$$X(s) = \frac{1}{2j} \left[ \frac{1}{(s - j2\pi)} - \frac{1}{(s + j2\pi)} \right]$$

$$= \frac{2\pi}{2j} \left[ \frac{1}{(s - j2\pi)(s + j2\pi)} \right]$$

$$= \frac{2\pi}{(s^2 + 4\pi^2)}$$

**Figure 3**: A sample problem from the experiment (left panel), pictured here with the detailed solution that was provided as feedback (right panel).

2.1.1.4 Procedure

Students completed weekly multiple-choice homework assignments using OpenStax Tutor. They were required to work on these assignments individually, and the solutions were due one week after the homework was assigned. Entering a solution involved two steps. First, students typed their solution in a free-form textbox or uploaded
an image file. Second, after submitting the free-form response, they were shown a set of multiple-choice alternatives and asked to select the correct answer. The purpose of this two-step procedure was to incorporate retrieval practice (i.e., practice generating a solution on one’s own, without the aid of the multiple-choice alternatives), while also enabling the automated scoring of the multiple-choice responses. The initial free-form responses were not graded.

Feedback on the homework problems was released either immediately after the assignment deadline or one week after the deadline. Students received an email notification when the feedback was available for viewing. To view the feedback, students clicked on a unique link to open an individual feedback screen for each problem. The feedback screen contained the initial problem, the correct multiple-choice answer, and a detailed solution. Students were required to view the feedback at least once in order to receive credit for each completed problem. After it had been released, the feedback remained available for repeated viewing for the duration of the semester.

Over the course of the semester, students took four exams (three unit exams and one final exam). Each exam contained new problems about the material that was covered in the homework assignments. Exams were closed book and were administered using the two-step procedure in the OpenStax Tutor system.
2.1.2 Results

Unless otherwise stated, all results were significant at the .05 alpha level. Two criteria had to be met for a student’s data to be included in the analyses: 1) the student had not taken the course before, and 2) the student had to have completed at least half of the homework assignments (4 or more of 8) and course exams (2 or more of 4). Two students were retaking the course and thus their data were excluded. All of the remaining 24 students met the criteria for inclusion in the analyses.

In addition, data from individual homework assignments and the corresponding exam problems were excluded from the analyses on a student-by-student basis according to the following criteria: 1) the student completed less than half of the problems for the homework assignment; 2) the student completed the assignment after the deadline and within 24 hours of the corresponding exam or after the exam. Applying these criteria resulted in the exclusion of the data related to 3% (6 out of 192) of homework assignments.

2.1.2.1 Homework Performance

There was no difference in the proportion of multiple-choice homework questions answered correctly as a function of whether students received immediate or delayed feedback [0.83 vs. 0.81; t < 1]. This result was not surprising because students’ responses were made before they received the feedback. It is important, however, because it
indicates a lack of baseline differences (i.e., in prior knowledge or inherent ability) between the two groups.

2.1.2.2 Feedback Viewing

Students were required to view the feedback; for the most part, they adhered to this guideline, loading the feedback screen for 98% of all homework problems. Students who received the feedback immediately after the assignment deadline were more likely to view it in comparison to students who received the feedback after a longer delay [99% vs. 94%; \( t(7) = 3.04, SED = 1.75, p = 0.02, d = 1.87 \)]\(^2\).

In the immediate feedback condition, an average of 4.1 days elapsed between when learners submitted the homework problems and when they viewed the feedback; in the delayed feedback condition, the average lag between homework completion and feedback was 11.6 days. Thus, students did not always view the feedback at precisely the time that it became available; as intended, however, the lag between homework completion and feedback viewing was approximately one week longer in the delayed feedback condition than the immediate feedback condition.

On average, students viewed the feedback for each problem 1.54 times (excluding the small percentage of problems for which students did not view feedback at all). Students who received immediate feedback viewed the feedback for a given problem a

\(^2\) Corrected for violation of the equality of variances assumption.
greater number of times than students who received delayed feedback, but this difference was not significant [1.65 vs. 1.32; \( t(22) = 1.66, SED = 0.20, p = 0.11, d = 0.72 \)].

2.1.2.3 Exam Performance

The critical dependent measure was performance on course exams. Students who received delayed feedback on the homework assignments answered a greater proportion of exam questions correctly than did students who received immediate feedback [.92 vs. .84; \( t(22) = 2.16, SED = 0.04, p = 0.04, d = 0.75 \)].

2.1.2.4 Student Impressions

See Appendix for the full survey results. Nineteen of the 24 students completed the final survey. Most of the students who received immediate feedback reported that they liked the timing of the feedback a lot (82%) and they thought that their performance in the course had benefited a lot from it (73%). In contrast, the majority of students who received delayed feedback reported that they did not like it (57%) and that their learning either did not benefit from it (43%) or was hurt by it (14%).

2.2 Experiment 2: The Role of Feedback Timing in Classroom Learning: A Within-Subjects Manipulation with Student Impressions

The primary objective of Experiment 2 was to replicate the major findings of Experiment 1. In Experiment 2, timing of feedback was manipulated within-subjects, allowing an examination of the generalizability of the effect of delaying feedback. A second advantage of switching to a within-subjects design is that all students in
Experiment 2 experienced both immediate and delayed feedback; thus, student perceptions could be drawn from a relative comparison between the two feedback timing conditions. The critical question was whether direct experience with both immediate and delayed feedback would help students to appreciate the benefits of delayed feedback for their learning.

Another goal of Experiment 2 was to determine how often students would view the feedback message when feedback viewing was not required, and whether delayed feedback would still incur an advantage under these circumstances. One possible hypothesis is that motivation to view the feedback declines over time; if this is the case, then students might be less likely to choose to view delayed feedback. Obtaining evidence for this hypothesis could help to explain the advantage of immediate feedback that emerged in some previous classroom studies. To examine this possibility, half of the students were required to view the feedback message in order to earn course credit in Experiment 2, but the other half were not.

2.2.1 Methods

2.2.1.1 Participants

Participants were students in two sections of the “Continuous Time Signals and Systems” course. Fifty students agreed to release their data for research purposes.
2.2.1.2 Design

The design was a 2 (timing of feedback: immediate, delayed) x 2 (feedback viewing: required, optional) mixed factorial. Timing of feedback was manipulated within-subjects, but between-materials; in other words, a student who received immediate feedback for the topics covered in Homework 1 would receive delayed feedback for the topics covered in Homework 2 (and vice versa). The feedback-viewing requirement was between-subjects; students in one section of the course were required to view feedback to earn credit, while students in the other section were not.

2.2.1.3 Materials

The materials were the same as in Experiment 2 except that the instructor gave only three exams. She also added one new topic and a corresponding homework assignment, resulting in a total of nine topics for which the homework and exam data were analyzed.

2.2.1.4 Procedure

The procedure was similar to Experiment 1 with a few exceptions. In each section of the class, students were split into two groups. Each group alternated from week to week between immediate and delayed feedback. Within each section, one group started with immediate feedback in the first week, whereas the other group started with delayed feedback. Thus, in any given week, some students in each section received immediate feedback while others received delayed feedback. As in Experiment 1, immediate
feedback was released immediately after the assignment deadline, whereas delayed feedback was released one week after the deadline. In addition, students in one section were required to view the feedback in order to get credit for the homework (as in Experiment 1), whereas students in the other section could choose whether or not to view feedback. Regardless of whether feedback viewing was required or optional, students received an email notification when the feedback was released.

2.2.2 Results

The criteria for inclusion were the same as in Experiment 2. Four students dropped the class before the exam and thus failed to contribute any useable data. Of the remaining 46 students who completed at least one exam, 9 students were excluded from the data analysis because they failed to complete at least half of all homework assignments (5 or more of 9) and / or exams (2 or more of 3). Most of the students who were excluded because of these criteria had dropped the course after the first exam. One additional student was excluded because the student registered in OpenStax Tutor for the wrong section of the course and thus did not receive the correct experimental conditions. After applying these criteria, a total of 36 students were included in the data analysis. We also excluded homework assignments (and their corresponding exam problems) according to same criteria as in Experiment 1. Applying these criteria resulted in the exclusion of the data related to 12% (38 out of 324) of homework assignments.
2.2.2.1 Homework Performance

The proportion of correct multiple-choice responses on the homework assignments was approximately equal across the two sections (overall $M = 0.70$), indicating that the two groups did not differ in knowledge prior to the implementation of the feedback manipulation. A 2 x 2 repeated measures ANOVA revealed no significant main effects of timing or feedback viewing requirement ($F$s < 1), and the interaction also was not significant [$F(1, 34) = 1.37, MSE = 0.01, \rho = 0.25, \eta^2 = 0.04$].

2.2.2.2 Feedback Viewing

When feedback viewing was required, students viewed the feedback for 94% of all problems (immediate feedback = 93%; delayed feedback = 95%). In contrast, when feedback viewing was optional, students only viewed the feedback for 47% of all problems (immediate feedback = 53%; delayed feedback = 45%). A 2 x 2 repeated measures ANOVA revealed a significant main effect of feedback-viewing requirement [$F(1, 34) = 30.38, MSE = 1186.78, \rho < 0.01, \eta^2 = 0.47$]; neither the main effect of feedback timing ($F < 1$) nor the interaction were significant [$F(1, 34) = 2.52, MSE = 194.03, \rho = 0.12, \eta^2 = 0.07$]. When feedback viewing was required, students viewed the feedback an average of 5.8 days after completing the corresponding homework assignment in the immediate feedback condition and 14.3 days in the delayed feedback condition. When feedback viewing was optional, the corresponding averages were 13.0 days and 20.1 days, respectively.
Figure 4 depicts the average number of times that students viewed the feedback for a given homework problem as a function of timing of feedback and feedback viewing requirement (excluding problems for which students did not view feedback at all). When students viewed the feedback, they did so a greater number of times if it was released immediately after the assignment deadline and they were required to view it. A 2 x 2 repeated measures ANOVA confirmed this observation by showing significant main effects of feedback timing \[ F(1, 34) = 18.28, \text{MSE} = 0.14, p < .01, \eta^2 = 0.35 \] and feedback-viewing requirement \[ F(1, 34) = 14.69, \text{MSE} = 2.06, p = 0.001, \eta^2 = 0.30 \]; however, the interaction was not significant \( F < 1 \).
2.2.2.3 Exam Performance

Figure 5 shows the proportion of correct multiple-choice responses on the exams as a function of timing of feedback and feedback-viewing requirement. Replicating the results of Experiment 1, students performed better on the exam when they had learned the material with delayed feedback. In addition, requiring students to view feedback also improved performance. A 2 x 2 repeated measures ANOVA revealed significant main effects of timing of feedback [$F(1, 34) = 4.79, MSE = 0.05, p = 0.04, \eta^2 = 0.12]$ and
feedback-viewing requirement \([F(1, 34) = 7.85, MSE = 0.11, p = 0.01, \eta^2 = 0.19]\).

Although the interaction was not significant \((F < 1)\), it is interesting to note that the size of the timing of feedback effect when feedback viewing was required \((d = 0.57)\) was about twice that when feedback viewing was optional \((d = 0.23)\).

![Figure 5](image)

**Figure 5:** Proportion of correct multiple-choice responses on the course exams as a function of timing of feedback and feedback viewing condition. Error bars indicate 95% CI.

### 2.2.2.4 Exam Performance as a Function of Retention Interval

One additional analysis was conducted to rule out an alternative explanation for the benefits of delaying feedback. Specifically, this analysis addressed the concern that the viewing of the delayed feedback took place closer in time to exams than the viewing
of the immediate feedback. One could speculate that the apparent benefit of delayed feedback was actually just a reflection of this difference in retention interval (see Metcalfe, Kornell, & Finn, 2009).

To examine this possibility, the data from the required feedback condition were re-analyzed as a function of the length of the retention interval from the assignment deadline to the exam. The retention interval was classified as “short” when the homework assignment was due 1 or 2 weeks before the corresponding exam and “long” when the homework deadline was 3 or 4 weeks beforehand. If the benefits of delaying feedback were in fact due to differing length retention intervals between the immediate and delayed feedback conditions, then the advantage of delayed feedback should be smaller at the long retention interval. As shown in Table 1, however, the benefit of delaying feedback was approximately equal at the short and long retention intervals. This impression was statistically confirmed in a 2 (feedback timing: immediate, delayed) x 2 (retention interval: short, long) ANOVA, which revealed a significant main effect of timing of feedback \(F(1, 15) = 5.56, MSE = 0.06, p = 0.03, \eta^2 = 0.12\), but no main effect of retention interval and no feedback-timing x retention interval interaction \((Fs < 1)\). In summary, there was no evidence that a shorter retention interval accounted for the benefits of delaying feedback. Note that there was inadequate power to conduct the same analysis using the data from Experiment 1, but the data show a similar pattern (Table 1).
Table 1: Average proportion of correct multiple-choice responses on the course exams in Experiments 1 and 2 as a function of short (1-2 weeks) versus long (3-4 weeks) retention interval between the homework assignment and corresponding exam. The means for Experiment 2 only represent performance in the required feedback viewing condition (the data from the optional feedback viewing condition were excluded from this analysis).

<table>
<thead>
<tr>
<th></th>
<th>Experiment 1</th>
<th></th>
<th>Experiment 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate</td>
<td>Delayed</td>
<td>Immediate</td>
<td>Delayed</td>
</tr>
<tr>
<td>Short (1-2 weeks)</td>
<td>.84</td>
<td>.92</td>
<td>.67</td>
<td>.81</td>
</tr>
<tr>
<td>Long (3-4 weeks)</td>
<td>.83</td>
<td>.94</td>
<td>.63</td>
<td>.81</td>
</tr>
</tbody>
</table>

2.2.2.5 Student Impressions

Twenty-nine of the 36 students filled out the final survey. Collapsing across section, almost all of the students indicated that they “highly preferred” immediate feedback (90%) and no students preferred delayed feedback. In terms of their performance in the course, most students reported that their performance “benefited a lot more” from immediate feedback (79%).

2.3 Discussion

In contrast to the common recommendation that feedback in the classroom should always be provided immediately, Experiments 1 and 2 showed that delaying feedback on homework assignments enhanced exam performance in a college engineering course. Interestingly, students were unaware of these benefits, reporting that they preferred immediate feedback and believed they had learned more from it. This metacognitive
disconnect occurred even when students had direct experience with both feedback-timing conditions (Experiment 2). These findings contribute to a growing literature showing that delaying feedback can improve learning, even in “real-world” settings outside of the laboratory. In the present experiments, students engaged in a variety of other learning activities (lectures, meeting with teaching assistants, reading the textbook, etc.) outside of the online homework assignments. There was no attempt to control the “noisiness” of the classroom environment, and yet the feedback-timing manipulation still affected long-term retention and transfer of course material.

As mentioned previously, there are two different (and not mutually exclusive) theoretical accounts that seek to explain the benefits of delaying feedback for later memory. According to the interference perseveration hypothesis, delaying feedback is beneficial because it allows learners to forget their errors, thereby lessening the interference that occurs when the correct answers are presented in the feedback message (Kulhavy & Anderson, 1972). Although it is possible that forgetting of errors accounted for some portion of the benefits of delaying feedback in Experiments 1 and 2, it is unlikely that this is the primary explanation for the results. As noted above, forgetting of errors is not correlated with better memory for the corrected information (Butler, Fazio, & Marsh, 2011). Moreover, in the present experiments, the exams consisted of transfer problems (i.e., problems that were different from those on the homework); thus, even if
specific errors from the homework problems were remembered, it seems unlikely that they would represent a significant source of interference.

The spacing hypothesis provides a more likely account for the present data. According to this hypothesis, delaying feedback benefits memory because it provides opportunities for distributed practice of items that were initially answered correctly. In the present experiments, learners performed quite well on the initial homework assignments \( (M = .82 \) correct in Experiment 1 and \( M = .70 \) correct in Experiment 2). Given this high level of initial performance, there were relatively few errors to correct, and the major function of the feedback was to reinforce learners’ correct understanding. In line with many findings from the broader literature (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006 for a review), the present results indicate that spaced review of the critical information results in larger gains for long-term memory.

Although the present study clearly demonstrates that delaying feedback does not impair student learning, the implications for the effectiveness of truly “immediate” feedback are less clear. In the present experiments, the immediate feedback was released to students after the assignment deadline rather than immediately after each problem. The choice of this particular timing for immediate feedback was driven by the need to protect against student cheating, which is likely to be a concern in many classrooms. Nevertheless, future research should explore the relative effectiveness of feedback that is given immediately after the completion of a problem to feedback given at various delays.
It is likely that there are some circumstances in which immediate feedback is at least as beneficial as delayed feedback. For example, if learners are less likely to view the feedback message after a delay (as suggested in our research), then immediate feedback may produce superior learning. It may also be the case that particular types of errors are best corrected immediately; for example, if an error is relatively subtle and difficult to notice, it may be best corrected when followed immediately by a feedback message. The next chapter of this dissertation examines this possibility.
3. Experiments 3 and 4: Correcting Strongly-Held Errors: A Case for Immediate Feedback

Upon receiving corrective feedback, learners are often quite good at correcting their mistakes, whether they were learning definitions of English vocabulary (Metcalf & Kornell, 2007) or science concepts such as the respiratory system (Butler, Godbole, & Marsh, 2013), brain regions (Lantz & Swawiski, 2014), or the solar system (Little & Bjork, 2014). Some errors, however, are not easily corrected. In particular, people often misremember the details of events or even falsely remember entire events that never occurred. It is notoriously difficult to avoid and correct such false memories. For example, hearing a list of semantically related words like bed, rest, and tired yields later claims that a non-presented word, sleep, was also on the list (the Deese-Roediger-McDermott [DRM] illusion; Roediger & McDermott, 1995). People misremember sentences like The new baby stayed awake all night as The new baby cried all night (Brewer, 1977). Answering leading questions like “How fast were the cars going when they smashed into each other?” evokes memories of (nonexistent) broken glass at the scene of an accident (Loftus & Palmer, 1974). Compared to other memory errors, false memories are often associated with vivid (but inaccurate) experiences of remembering, or the feeling that one recollects specific details of the event (Roediger & McDermott, 1995; Chan & McDermott, 2006). Two common attempts to correct these errors involve

---

specifically warning participants that an activity can yield false memories (e.g., Gallo, Roberts, & Seamon, 1997; McDermott & Roediger, 1998) and allowing multiple encodings of the to-be-remembered information before a memory test is given (e.g., McDermott & Chan, 2006; Watson, McDermott, & Balota, 2004). Unfortunately, neither method is particularly effective. A strong warning combined with a practice list and a full explanation of the DRM illusion still results in false recognition of nearly half of the critical lures (Gallo et al., 1997), and false recall of almost one-third of them (Watson et al., 2004). After three encodings of pragmatic inferences (e.g., *The new baby stayed awake all night*), learners still “recognize” the inference (false memory) answer on 28% of final test trials (McDermott & Chan, 2006).

Why is it so difficult to correct false memories, when it appears relatively simple to correct many other types of errors? One argument is that learners must first notice their errors in order to correct them. This requirement is simple in many cases, such as when the learner is aware that he or she has no idea of the answer (e.g., you likely know if you don’t know the translation of the Luganda word *leero*). However, almost by definition a false memory means that learners are unaware of their mistakes—such memories are accompanied by confidence and the subjective (but false) experience of recalling sounds, feelings, or other experiences from the original event (Roediger & McDermott, 1995; Chan & McDermott, 2006). The vividness of these errors may make the learner resistant to feedback, similar to the case in which two people both claim a memory as their own,
despite knowing that the event could have only happened to one of them (disputed memories; Sheen, Kemp, & Rubin, 2001). In addition, most feedback about false memories is not as explicit as someone else telling you that a memory is theirs (and not yours). One of the most common approaches is to give the learner multiple study-test trials; however, success requires noticing one’s intrusion was not actually on the list (e.g., Kensinger & Schacter, 1999; McDermott, 1996) or in the passage (Fritz, Morris, Bjork, Gelman, & Wickens, 2000; Kay, 1955). Learners establish a schema for the event, making it difficult to notice and correct memories that are schema-consistent.

The notion that learners may fail to notice their false memories, even when confronted again with the correct information, leads to a prediction for how best to correct them: a successful correction procedure must draw attention to learners’ mistakes. There are several ways to accomplish this; perhaps the most straightforward approach is to present corrective feedback immediately after each error is committed (i.e., on a trial-by-trial basis). Essentially, this is equivalent to telling the learner “no, sleep wasn’t on the list; it was bed” as soon as sleep is falsely recalled. Note that this prediction—that immediate feedback should best facilitate the correction of false memories—is in contrast to other findings in the broader feedback literature, in which feedback administered after a brief delay often yields improved performance, presumably because delayed feedback serves as a spaced study trial (Butler, Karpicke, & Roediger, 2007; see Pashler, Rohrer, Cepeda, & Carpenter, 2007 for a review of the benefits of spacing.
practice over time). Importantly, however, these studies did not involve false memories; instead, they corrected errors in general knowledge (Smith & Kimball, 2010); history (Butler & Roediger, 2008) and engineering concepts (Experiments 1 and 2 of this dissertation), among other educationally-relevant materials. Experiments 3 and 4 evaluate the benefits of immediate versus delayed feedback for correcting false memories.

Of course, the provision of immediate, trial-by-trial feedback is not the only way to draw learners’ attention to their errors. The purpose of Experiment 4 is to examine a second way to accomplish the same goal: explicitly asking learners to evaluate their past responses at the time that delayed feedback is presented (i.e., by asking, Was your [previous] answer correct?). Regardless of the specific procedural details, any situation that encourages learners to notice the discrepancies between their errors and the correct information should result in a reduction of false memories.

In both experiments, the key hypotheses were tested using pragmatic inference materials whereby sentences such as The karate champion hit the cinderblock are misremembered as The karate champion broke the cinderblock. These inferences are known to produce phenomenological experiences “indistinguishable from those of true memories” (p. 633, Chan & McDermott, 2006), including high confidence in one’s wrong responses (Sampaio & Brewer, 2009).
3.1 Experiment 3

Many studies with educational materials have shown that long-term retention is enhanced when learners receive feedback after a delay (e.g., Anderson, Kulhavy, & Andre, 1972; Butler & Roediger, 2008; Carpenter & Vul, 2011; Metcalfe, Kornell, & Finn, 2009; Phye & Andre, 1981; Sassenrath & Yonge, 1968; Smith & Kimball, 2010). Delaying feedback is not always beneficial, however. In particular, this advantage disappears when correcting high-confidence errors in general knowledge (e.g., Sydney is the capital of Australia; Vitamin C cures colds; Sitzman, Rhodes, & Tauber, 2014). In the case of these high-confidence errors, delaying feedback likely reduces the chance of the learner noticing the contradiction between the feedback and their prior mistake.

Experiment 3 examined the possibility that false memories are another class of errors that do not benefit from delaying feedback. As described earlier, false memories differ from most errors in that they are vivid, held with confidence, and involve thinking back to a particular time and place—factors that likely make it difficult to notice the contradiction between one’s error and delayed feedback. The purpose of Experiment 3 was to test the prediction that immediate, trial-by-trial feedback is more effective for correcting false memories, because it enables a back-to-back comparison between the correct response and one’s error.
3.1.1. Methods

3.1.1.1 Participants

Seventy-eight Duke University undergraduates participated in exchange for course credit ($n = 26$ for the immediate and $n = 26$ for each of two delayed feedback conditions; see Procedure for full explanation of these groups).

3.1.1.2 Design

Participants received immediate feedback or one of two schedules of delayed feedback.

3.1.1.3 Materials

We conducted pilot testing on McDermott and Chan’s (2006) pragmatic inference materials and identified 30 items (e.g., *The ugly stepsisters asked Cinderella to mop the floor*; pragmatic inference answer: *told*) for use in our experiment (the critical word(s) were not underlined in the versions that were shown to participants). For many of the sentences, only one word was required to fill in the blank; for other sentences, a few words were required ($M = 1.80$ words).

We also created 12 filler sentences that did not contain pragmatic implications (e.g., *The boy slipped on the banana peel*), to make the task similar in length to McDermott and Chan (2006). For each critical and filler item, we created a sentence fragment (e.g., *The ugly stepsisters ____ Cinderella to mop the floor*) to use on the initial and final tests.
3.1.1.4 Procedure

The experiment was programmed with E-prime 2.0 software. During study, participants were instructed to read and remember the sentences. They read the 42 sentences (30 critical sentences and 12 fillers) at a rate of 4 s each, with a 500-ms blank screen and a 1-s fixation point between trials. Participants then solved unrelated brainteasers for 10 min.

Next, participants completed the self-paced initial test. For each sentence fragment, they were instructed to fill in the missing word(s), being careful to use the exact wording from the sentence that they had studied. Participants were told not to guess, and were asked to enter “I don’t know,” if they could not remember the critical word(s).

For all participants, the feedback took the form of a 4-s re-presentation of the originally studied (correct) sentence (i.e., correct answer feedback). See Figure 6 for a schematic of the design. For the immediate feedback condition, the feedback was presented immediately after each initial test trial. Participants in the immediate feedback condition then solved unrelated brainteasers for 15 min before beginning the final test. The delayed feedback was administered according to one of two schedules. In both schedules, participants completed the initial test (without viewing any feedback), solved unrelated brainteasers for 5 min, and then received the feedback, which was essentially another chance to see the study list. What differed across the schedules was the lag
between the feedback presentation and the final test (10 min of brainteasers in the first schedule and 15 min of brainteasers in the second schedule). Thus, one delayed feedback condition equated the lag between the initial and final tests across feedback groups. In the other delayed feedback condition, the lag between the presentation of the feedback and final test was equated. Both schedules were necessary to ensure that a possible advantage of delaying feedback was not due to the delayed feedback being presented closer in time to the final test (Metcalfe, Kornell, & Finn, 2009).

Figure 6: A schematic of the design for the immediate feedback (IFB) condition and the two delayed feedback (DFB) conditions in Experiments 3 and 4. The first delayed feedback condition controls for the lag between study and final test, whereas the second delayed feedback condition controls for the lag between feedback and the final test.
Participants then completed the final test, which was exactly the same as the initial test except that none of the participants received feedback.

3.1.2 Results

3.1.2.1 Data Scoring and Analysis

Participants’ responses were coded as correct, inference, “I don’t know,” or another wrong answer. Consistent with past research (McDermott & Chan, 2006), we identified a list of *a priori* responses that would be defined as correct or inference answers. For example, for the sentence *The ugly stepsisters asked Cinderella to mop the floor, asked* was classified as correct, and *told, ordered, and forced* were classified as inferences. Other responses that had not been defined *a priori* were coded as other wrong answers. Two independent coders scored the responses (Cohen’s kappa = .95), and a third coder resolved discrepancies. The results are presented in Table 2; note that changes in one response category across tests necessarily produce changes in the other categories (e.g., an increase in correct responses necessarily coincides with a decrease in incorrect answers). Because the two delayed feedback schedules resulted in virtually identical performance on both the initial and final tests, we collapsed across them to form one delayed feedback condition for the reporting of the statistical analyses reported here (but the interested reader can find the complete breakdown of means in Table 2). To examine the relative effectiveness of immediate and delayed feedback, we examined the
proportion of final test items completed correctly versus with the critical inferences in separate 2 (test: initial, final) x 2 (feedback timing: immediate, delayed) ANOVAs.

3.1.2.2 Correct Answers

Participants who received immediate feedback produced more correct answers initially ($M = .24$) than those who received delayed feedback ($M = .19$), $F(1, 76) = 6.83, p = .01, MSE = .033, \eta^2 = .001$. Although the advantage of the immediate feedback condition became numerically larger on the final test ($M = .79$ for the immediate vs. .68 for the delayed feedback condition), the test x feedback timing condition interaction was not significant, $F(1, 76) = 2.22, p = .14, MSE = .01, \eta^2 = .003$.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Inference</th>
<th>Other Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>Final</td>
<td>Initial</td>
</tr>
<tr>
<td>Immediate</td>
<td>.24 (.13)</td>
<td>.79 (.15)</td>
<td>.34 (.12)</td>
</tr>
<tr>
<td>Delayed 1</td>
<td>.21 (.12)</td>
<td>.69 (.20)</td>
<td>.33 (.13)</td>
</tr>
<tr>
<td>Delayed 2</td>
<td>.18 (.14)</td>
<td>.67 (.15)</td>
<td>.33 (.12)</td>
</tr>
</tbody>
</table>

3.1.2.3 Inferences

Our primary interest was in the intrusion of inferences when recalling the original sentences. Initially, participants produced the critical inferences on about one-third of
trials, and this rate did not differ across the immediate ($M = .34$) and delayed ($M = .33$) feedback conditions. Feedback was very helpful; overall, the proportion of sentence fragments completed with the critical errors was reduced close to floor levels ($M = .07$) on the final test. However, there was still evidence that feedback timing mattered: participants in the delayed feedback conditions used the inferences to complete 9% of final sentence fragments, whereas participants receiving immediate feedback only completed fragments with the inferences 4% of the time [$F(1, 76) = 4.67, p = .03, MSE = .007, \eta^2 = .01$ for the test x feedback timing interaction]. This pattern emerged even though the timing of feedback manipulation was relatively subtle, with delayed feedback administered only 5 min later than immediate feedback. Both immediate and delayed feedback greatly reduced the proportion of false memories produced across tests, but immediate feedback was more effective.

### 3.2 Experiment 4

As predicted—but in contrast to patterns often observed with other types of errors—immediate feedback was superior to delayed feedback in reducing the number of false memories. In addition to replicating Experiment 3, the central goal of Experiment 4 was to more clearly evaluate the reason for the advantage of immediate feedback, as well as to examine whether performance under delayed feedback could be improved to the same level. Specifically, we tested the idea that delayed feedback might be just as effective as immediate feedback if learners were explicitly required to compare the
feedback messages to their prior responses. Broadly speaking, previous research has already shown that noticing discrepancies is critical for error correction. For example, participants who are told to replace their memories of a previously-studied cue-target word pair (e.g., knee-bone) with an updated pair (e.g., knee-bend) are more successful at doing so if, at the time of encoding the second pair, they notice that the target word has changed (i.e., notice the discrepancy between the two pairs; Walheim & Jacoby, 2013). Moreover, a comparison of the data from Walheim and Jacoby (2013)’s Experiments 1 and 2 suggests that an explicit requirement to look for changes across trials may increase the likelihood of noticing such discrepancies. In line with this idea, in Experiment 4 we manipulated whether learners were explicitly prompted to compare each feedback message to their (past) response (thereby helping them to notice to any discrepancies). Directly after receiving the immediate or delayed feedback, half of the learners were required to answer the question “Was your [previous] answer correct?”—a judgment that should encourage them to bring their previous response to mind while viewing the feedback. This design allowed us to replicate the surprising benefit of immediate feedback from Experiment 3, as well as to examine whether another manipulation could successfully promote comparisons between one’s errors and the correct information.

3.2.1 Methods

3.2.1.1 Participants

Participants were workers on Amazon Mechanical Turk (MTurk), an online
marketplace where people complete tasks in exchange for payment. Our records from Qualtrics survey software—which was used to present the experiment—showed that 315 workers clicked on the experiment link, with many of them ultimately deciding not to complete the experiment (of those who quit, the vast majority [86%] did so when they had progressed through less than 30% of the survey). We continued collecting data on MTurk until we had reached our goal of 30 participants in each of six conditions who had completed the full experiment (180 participants in total).

3.2.1.2 Design

The design was a 3 (feedback timing: immediate, delayed schedule 1, delayed schedule 2) x 2 (presence of follow-up question: yes, no) mixed factorial.

3.2.1.3 Materials

Materials were the same as in Experiment 3, but the experiment was presented using Qualtrics survey software.

3.2.1.4 Procedure

As in Experiment 3, participants received either immediate or delayed feedback, and delayed feedback was administered according to one of two schedules (lag of 10 min between the feedback and the final test vs. lag of 15 min between the feedback and the final test; see Figure 6). Unlike Experiment 3, half of the learners were asked, upon receiving either the immediate or delayed feedback, “Was your [previous] answer correct?” (with the option to select either “Yes” or “No”). This question was meant to
ensure that learners would make a direction comparison between their own answers and the feedback message.

3.2.2 Results

3.2.2.1 Data Scoring and Analysis

Scoring was the same as in Experiment 3 (Cohen’s kappa = .96); the means are shown in Table 3. The proportion of correct and inference answers were included in separate 2 (test: initial, final) x 2 (feedback timing: immediate, delayed) x 2 (presence of follow-up question: yes, no) ANOVAs. As a manipulation check, we note that participants in all three conditions with the follow-up question were very good at judging whether the feedback matched their answers ($M = .91$).

3.2.2.2 Correct Answers

Overall, participants correctly answered about 19% of initial test trials; after receiving feedback, correct responding increased dramatically ($M = .57$ on the final test). There was no main effect of feedback timing, $F(1, 176) = 1.07, p = .30, MSE = .09, \eta^2 = .004$, and no main effect of the presence of the follow-up question, $F < 1$. There was, however, a significant test x follow-up question interaction, $F(1, 176) = 5.23, p = .02, MSE = .02, \eta^2 = .001$; the increase in correct responding across tests was greater when participants were required to answer the follow-up question (increasing from .17 to .59) than when they were not (a smaller increase, from .21 to .55).
3.2.2.3 Inferences

Again, we were most interested in the production and subsequent reduction in the rate of false memories. As can be seen in Table 3, there was some baseline variability across conditions, likely due to the less controlled conditions involved when testing MTurk subjects; what is more important is the reduction in the rate of false memories across tests. Replicating Experiment 3, there was a significant test x feedback timing interaction, $F(1, 176) = 12.83, p < .001, \text{MSE} = .012, \eta^2 = .06$, with the reduction in errors across tests depending on whether participants had received immediate or delayed feedback. Whereas the inference rate decreased by 28 percentage points in the immediate feedback condition, it only dropped by 18 points after delayed feedback. These data replicate the main finding of Experiment 3.

Critically, there was also a significant test x follow-up question interaction, $F(1, 176) = 12.35, p = .001, \text{MSE} = .012, \eta^2 = .06$; the reduction in errors varied depending on whether learners had been required to answer the follow-up question (“Was your answer correct?”) when they received feedback. Learners produced slightly more errors when required to answer the follow-up question on the initial test ($M = .35$ versus $.33$ for the control), but the rate of errors dropped to only 10% when subjects were explicitly prompted to compare their responses to the feedback, as opposed to a final inference rate of 15% for the control.

Although the three-way interaction of test, feedback timing, and presence of a
follow-up question did not reach significance ($F(1, 176) = 2.00, p = .16, \text{MSE} = .012, \eta^2 = .01$), Table 3 suggests that reminding subjects to explicitly compare the feedback to their responses was effective at mitigating the negative effects of delayed feedback. To confirm this impression, we conducted additional analyses on the data from the delayed feedback conditions. We directly compared the proportion of questions answered with inferences as a function of whether or not participants had received the follow-up question. There was no difference in the initial rate of inferences as a function of whether the delayed feedback was paired with the follow-up question ($t < 1$); however, participants who received delayed feedback with the follow-up question responded with fewer inferences on the final test ($M = .10$), compared to those who received delayed feedback without the follow-up question ($M = .18$), $t(118) = 3.16, p = .002, d = 0.58$.

The addition of the follow-up question—which encouraged a direct comparison between one’s errors and the correct answers—helped learners to avoid reproducing their mistakes, even when receiving delayed feedback.
Table 3: Proportion of sentence fragments answered correctly versus with inferences or other wrong answers in Experiment 4, as a function of (1) whether participants received immediate feedback or one of the two delayed feedback schedules and (2) whether the feedback message was paired with a follow-up question.

Note: Standard deviations are in parentheses.

<table>
<thead>
<tr>
<th>Follow-up Question</th>
<th>Correct</th>
<th>Inference</th>
<th>Other Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>Final</td>
<td>Initial</td>
</tr>
<tr>
<td>IFB</td>
<td>.13 (.13)</td>
<td>.59 (.21)</td>
<td>.44 (.17)</td>
</tr>
<tr>
<td>DFB 1</td>
<td>.23 (.30)</td>
<td>.55 (.31)</td>
<td>.27 (.16)</td>
</tr>
<tr>
<td>DFB 2</td>
<td>.17 (.19)</td>
<td>.61 (.24)</td>
<td>.34 (.15)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Follow-up Question</th>
<th>Correct</th>
<th>Inference</th>
<th>Other Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>Final</td>
<td>Initial</td>
</tr>
<tr>
<td>IFB</td>
<td>.25 (.25)</td>
<td>.61 (.21)</td>
<td>.32 (.19)</td>
</tr>
<tr>
<td>DFB 1</td>
<td>.15 (.11)</td>
<td>.48 (.30)</td>
<td>.33 (.16)</td>
</tr>
<tr>
<td>DFB 2</td>
<td>.19 (.22)</td>
<td>.50 (.28)</td>
<td>.35 (.17)</td>
</tr>
</tbody>
</table>

### 3.3 Discussion

Experiments 3 and 4 examined the correction of false memories, a type of error that is known for its stubbornness and persistence. The correction of these errors was greatly enhanced when the experimental conditions drew learners’ attention to the fact that they had made a mistake, either through the administration of trial-by-trial feedback
(Experiment 3) or through a specific prompt to compare their own previous answer to the feedback message (Experiment 4).

In the absence of these optimized correction procedures, why did learners sometimes fail to notice when they produced a false memory? In part, many false memories may go unnoticed because they are held with high confidence and vividness. The close semantic relationship between a false memory and the correct information likely also contributes to learners’ failures to notice their mistakes. That is, the false memory for *sleep* occurs precisely because this word shares a strong semantic relationship with the presented words; similarly, the false memory that *The new baby cried all night* is produced precisely because the correct version of the sentence was “designed to lead the listener to make schema-based inferences” (Sampaio & Brewer, 2009, p. 159). By definition, learners’ errors share close semantic overlap with the correct information, which likely makes these errors particularly difficult to notice. The literature on semantic illusions supports the same point: learners are surprisingly willing to provide meaningful answers to nonsensical questions like “How many animals of each kind did Moses take on the ark?”, as long as the errorful term (*Moses*) shares a close semantic overlap with the correct answer (*Noah*; Erikson & Mattson, 1981; van Oostendorp & de Mul, 1990).

In explaining the success of the current interventions, it should be noted that the correction procedures not only drew learners’ attention to the fact that they had made
errors, but also provided them with an opportunity to review the correct information. As demonstrated many times in the broader literature on error correction (Bangert-Drowns et al., 1991; Fazio et al., 2010; Shute, 2008), learners often reproduce the same mistakes unless they receive correct information to replace them. This same principle has also been demonstrated in the false memory literature (e.g., McConnell & Hunt, 2007; Lewandosky et al., 2012), and it is consistent with other findings that “forgetting” an old memory is more difficult than replacing it. For example, in the directed forgetting literature, participants first encode some items (e.g., a list of words), and are then asked to intentionally forget that material and to learn some new information (e.g., a different list of words). Importantly, the encoding of List 2 is critical in facilitating the forgetting of List 1. In other words, after receiving the “forget” cue, participants must receive competing material to encode; otherwise, they are unable to forget the first list (Gelfand & Bjork, 1985; Pastotter & Bauml, 2007, 2010).

The present results demonstrate that feedback can be successfully used to correct false memories, and that it is most effective when it is received directly after committing an error, or when the correction conditions otherwise facilitate a direct comparison between learners’ errors and the correct answers. Despite a growing number of studies that advocate for the efficacy of delayed feedback, false memories may represent a particular circumstance in which errors are best corrected immediately.
4. Experiment 5: Effects of Motivation on Feedback Processing and Learning from Feedback

Motivation is the general desire or willingness to initiate or persist in a goal-directed behavior. In general, there is a strong belief among students, educators, and researchers that greater motivation leads to better memory performance and academic achievement (e.g., Deci, Vallerand, Pelletier, & Ryan, 1991; Dweck, 1986; Lepper, Greene, & Nisbett, 1973; Ngaosuvan & Mantyla, 2005). There are several different mechanisms through which motivation is thought to influence memory. For example, one perspective assumes that the completion of academic tasks requires the use of self-control resources, which are limited and become depleted over time (e.g., Muraven & Baumeister, 2000). When learners are highly motivated—for example, by the promise of external reward or because they believe that the learning task is important—they are better able to compensate for the gradual fatiguing of their self-control “muscle” and continue performing at a high level, at least up to a certain point (Muraven & Slessareva, 2003). According to another perspective, motivation improves memory because it increases arousal and boosts attention to incoming stimuli (e.g., Seitz & Watanabe, 2005). Finally, from a biological point of view, motivation—and, in particular, the anticipation of reward—involves the release of dopamine, which promotes hippocampal memory consolidation (e.g., Lisman & Grace, 2005). Note that these explanations for the relationship between motivation and memory performance are not mutually exclusive. It
may be the case that motivation affects memory through all of these mechanisms, as well as through additional pathways.

The purpose of Experiment 5 was to examine the role of learner motivation in learning from feedback. Within the existing literature, a few recent studies address a similar goal. First, several teams of researchers have examined the possible benefits of curiosity for learning, with a specific interest in how curiosity affects feedback processing (e.g., Gruber, Gelman, & Raganath, 2014; Kang et al., 2009; Mullaney, Carpenter, Grotenhuis, & Burianek, 2014). The common finding coming out of this work is that learners tend to remember information better when they had previously stated that they were highly curious to learn it. Moreover, curiosity seems to benefit memory at least in part because it causes learners to process the feedback messages differently. For example, Gruber et al. (2014) found enhanced activity in the nucleus accumbens—a central player in the reward circuit of the brain—during anticipation of the feedback message for trivia facts that participants were highly curious to learn. In other words, a feeling of high curiosity is accompanied by an altered neural state at the time of feedback presentation, which may help to explain the benefits of curiosity for later memory.

In another set of experiments, Kang and Pashler (2014) attempted to directly manipulate learners’ motivation to learn a series of Swahili-English word pairs through the promise of incentives: a monetary bonus (Experiments 1 and 2) or the ability to leave the experiment early (Experiment 3). Interestingly, only the timing-savings manipulation
benefited later memory. However, this finding is difficult to interpret for at least two reasons. First, these studies did not include a manipulation check to confirm that the high-motivation conditions were truly motivating. It may be, for example, that participants were not particularly interested in earning a relatively small amount of money ($0.05-$0.30) for each correctly recalled item, thus explaining why the monetary incentive did not improve performance. Second, these experiments included experimenter-paced feedback, which might undermine the ability to find an effect of differing levels of motivation on later memory. In other words, it is possible that learners in the high-incentive condition would have chosen to spend more time studying the feedback (with the possibility of downstream benefits for memory performance), but the experimenter-controlled feedback meant that feedback study time was the same for everyone.

The goal of the present experiment was to contribute to the current effort to understand how motivation affects memory performance, with a specific interest in how motivation affects feedback processing. Motivation was manipulated through the task instructions, which were designed to either make the learning task seem highly important and relevant to participants’ lives (high-motivation instructions) or to make the task seem unimportant and irrelevant (low-motivation instructions). This instructional manipulation was chosen for two major reasons. First, previous research has shown that different framings of task importance can produce differences in self-rated motivation and performance (e.g., Muraven & Slessareva, 2003). Second, this particular manipulation
seemed likely to affect learners’ motivation levels without increasing their incentive to cheat, which was an important practical consideration because the experiment was conducted online. Overall, we predicted that heightening learners’ motivation would increase the amount of time that they spent studying the feedback messages, with benefits for later memory.

Another goal of Experiment 5 was to examine whether the influence of learner motivation on memory interacted with the effects of feedback timing. One concern about administering delayed feedback in the classroom and other applied settings is that learners’ motivation to view the feedback messages may decline over time. In other words, learners may no longer be interested in learning the correct answers when feedback is administered after a lengthy delay (see Experiment 2 of this dissertation for related evidence). If this is indeed the case, then delaying feedback may not be practically worthwhile; potential benefits of delaying feedback may be counteracted by learners’ unwillingness to view it. With this possibility in mind, one objective of Experiment 5 was to evaluate how the instructional manipulation, designed to alter learners’ motivation, would affect memory performance under immediate versus delayed feedback. Specifically, we suspected that the high-motivation instructions might have a particularly pronounced benefit when feedback was delayed, because they would prevent a decline in willingness to view the feedback that might otherwise occur.
4.1 Methods

4.1.1 Participants

Participants were 186 workers on Amazon’s Mechanical Turk. We excluded the data from 30 workers: 27 who admitted to cheating on a post-experimental questionnaire\(^1\); 1 who did not admit to cheating but whose initial test score was more than 3 standard deviations above the mean, suggesting that he/she likely cheated; and 2 who reported having prior knowledge of the Swahili language. The final sample contained 155 participants (n = 26 in five out of the six conditions and n = 25 in one of them).

4.1.2 Design

The design was a 2 (motivational instructions: high motivation, low motivation) x 3 (feedback timing: immediate feedback, delayed feedback schedule 1, delayed feedback schedule 2) between-subjects factorial. As in Experiments 3 and 4, delayed feedback was administered according to one of two schedules, in order to have one schedule that controlled for the lag from the initial to the final test across feedback groups and another schedule that controlled for the lag from the feedback presentation to the final test (see Figure 7). Because the two delayed feedback schedules did not always produce the same results (see Results section), we did not collapse across them.

---

\(^1\) Seventeen of these learners had received high-motivation instructions and 10 had received low-motivation instructions.
4.1.3 Materials

The learning materials were 34 Swahili-English word pairs (e.g., kasuku—parrot) selected from norms prepared by Nelson and Dunlosky (1994).

4.1.4 Procedure

The experiment was presented using Qualtrics survey software. See Figure 7 for a schematic of the procedure. First, participants studied the 34 critical word pairs at a rate of 4 s each.

Next, they completed a self-paced initial test; on each trial, a Swahili word was presented and participants were asked to type the English equivalent.

The manipulation of the motivational instructions occurred after the initial test. The manipulation was introduced at this point rather than earlier in the procedure because we were specifically interested in how motivation levels might affect feedback processing; if the manipulation had occurred earlier, it might have also affected initial study behavior. Participants were told that they would receive another opportunity to review the word pairs, and that they should try to learn them so that they would be able to remember them again later. Critically, the high-motivation instructions emphasized the potential relevance and importance of learning these materials by stating that:

Research has shown that performance on simple memory tests often predicts performance on other tasks, including academic performance, standardized test scores, and mentally-involving games like chess. Thus, performance on this task may tell you something about your broader cognitive abilities.

In contrast, the low-motivation instructions stated that:
Research has shown that performance on simple memory tests does not necessarily predict performance on other tasks. Thus, performance on this task may not tell you anything about your broader cognitive abilities.

After reading either the high- or low-motivation instructions, participants completed a manipulation check by answering the following questions using a 1-5 Likert scale: “Will performance on this task reveal anything about your broader cognitive abilities?”, “How motivated are you to learn the Swahili-English word pairs?” and “How important is it for you to perform well on this task?” A rating of 5 indicated the highest agreement.

Next, participants in the immediate feedback condition began self-paced review of the word pairs. In both of the delayed feedback conditions, participants completed a 10-min filler task (math problems) and then reviewed the feedback.
Figure 7: A schematic of the procedure for Experiment 5.

After the presentation of the feedback, participants completed a filler task of math problems for 10 minutes (delayed feedback 1 schedule) or 20 minutes (immediate feedback and delayed feedback 2 schedules). Finally, participants took the final test, which was exactly the same as the initial test.

4.2 Results

4.2.1 Manipulation Check: Ratings of Predictive Power of the Task, Learner Motivation, and Task Importance

As expected, learners were more likely to agree that performance in the experiment was predictive of their broader cognitive abilities after reading the high-motivation ($M = 3.83, SD = .97$) as opposed to the low-motivation ($M = 2.16, SD = .83$)
instructions, $t(154) = 11.55, p < .001, SED = .15, d = 1.85$. Learners who received the high-motivation instructions also reported being more motivated to learn the word pairs ($M = 4.14, SD = .95$) than those in the low-motivation condition ($M = 3.75, SD = 1.2$), $t(154) = 2.31, p = .02, SED = .17, d = .36$, but they did not rate the experimental task as being significantly more important [$M_{\text{high-motivation}} = 4.00, SD = 1.04$ vs. $M_{\text{low-motivation}} = 3.74, SD = 1.09, t(154) = 1.44, p = .15, SED = .17, d = .24$]. Overall, the high-motivation instructions were successful in affecting learners’ motivation to learn the word pairs.

4.2.2 Initial Test Performance

Overall, the initial test was difficult ($M$ proportion correct = .23; see Table 4).

There were no differences in performance as a function of motivation instructions or feedback-timing condition (both $F$s < 1), nor was the interaction significant, $F(2, 149) = 2.66, MSE = .05, p = .07, \eta^2 = 0.03$. Note that the absence of significant effects was expected, as neither the feedback-timing manipulation nor the motivation instructions had occurred at the time of the initial test.

Table 4: Proportion of Swahili-English word pairs correctly remembered on the initial test, presented separately for each of the six between-subjects conditions.

<table>
<thead>
<tr>
<th></th>
<th>Immediate</th>
<th>Delayed 1</th>
<th>Delayed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low motivation</td>
<td>.29 (.22)</td>
<td>.23 (.18)</td>
<td>.15 (.21)</td>
</tr>
<tr>
<td>High motivation</td>
<td>.23 (.22)</td>
<td>.22 (.21)</td>
<td>.28 (.24)</td>
</tr>
</tbody>
</table>

Note: Standard deviations are in parentheses.
4.2.3 Feedback Viewing Time

Did learners who received the high-motivation instructions spend more time viewing the corrective feedback? In examining this question, the data were trimmed to remove outliers, or response times that were more than two standard deviations above or below each participant’s individual mean. Note that the outcome of the analysis was the same when the untrimmed response times were included. Learners viewed each feedback screen for an average of 4.69 seconds; this value did not vary according to whether participants had received high- (\(M = 4.36\) seconds, \(SD = 3.59\)) versus low-motivation (\(M = 5.02\), \(SD = 4.82\)) instructions (\(F < 1;\) Table 5). Amount of time spent viewing the feedback also did not depend on whether learners had received the feedback immediately or according to one of the delayed schedules [\(F(2, 149) = 1.94, MSE = 17.99, p = .15, \eta^2 = 0.03\)], and there was no interaction between motivational-instructions and feedback-timing procedure (\(F < 1\)). Although learners who received the highly motivating instructions rated themselves as being more motivated to learn the word pairs, this did not translate into more time spent viewing the corrective feedback. Note, however, that these data were collected using a Qualtrics program on MTurk, which is not the ideal approach for gathering response time data (e.g., Simcox & Fiez, 2014).
Table 5: Average time spent viewing the feedback message, presented separately for each of the six between-subjects conditions.

Note: Standard deviations are in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Immediate</th>
<th>Delayed 1</th>
<th>Delayed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low motivation</td>
<td>4.32 (3.39)</td>
<td>5.85 (5.02)</td>
<td>4.87 (5.77)</td>
</tr>
<tr>
<td>High motivation</td>
<td>4.07 (3.18)</td>
<td>5.42 (4.09)</td>
<td>3.60 (3.29)</td>
</tr>
<tr>
<td>M</td>
<td>4.19 (3.25)</td>
<td>5.64 (4.54)</td>
<td>4.23 (4.69)</td>
</tr>
</tbody>
</table>

4.2.4 Final Test Performance

In contrast to initial predictions, motivational instructions had no effect on final test performance ($F < 1$); that is, participants were no better at learning the word pairs when they were highly motivated ($M = .43, SD = .31$) than when they were less motivated ($M = .42, SD = .28$; see Table 6). When considered in combination with the results of the preceding feedback-timing analysis, this makes sense; although learners who received the high-motivation instructions reported being more motivated to learn the word pairs, they may not have put any more effort into studying the feedback.

The timing of feedback manipulation did affect final test performance, however $[F(2, 149) = 3.98, MSE = .08, p = .02, \eta^2 = 0.05 ]$. A Tukey’s HSD post-hoc test revealed that final memory performance differed between participants who received immediate feedback ($M = .49, SD = .32$) and those who received the second delayed feedback schedule ($M = .34, SD = .27$); however, neither of these groups differed
significantly from participants who received the first delayed feedback schedule ($M = .45, SD = .27$). This pattern does not support the prediction that delaying feedback benefits memory; instead, it provides weak evidence that immediate feedback is beneficial.

There was no evidence of an interaction between motivational instructions and feedback timing ($F < 1$). The motivational instructions affected self-report ratings regarding the predictive power of the task and motivation to perform well, but they did not affect final memory performance, nor did they modulate the effectiveness of immediate vs. delayed feedback.

**Table 6: Proportion of Swahili-English word pairs correctly remembered on the final test, presented separately for each of the six between-subjects conditions.**

<table>
<thead>
<tr>
<th></th>
<th>Immediate</th>
<th>Delayed 1</th>
<th>Delayed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low motivation</td>
<td>.52 (.27)</td>
<td>.45 (.26)</td>
<td>.29 (.27)</td>
</tr>
<tr>
<td>High motivation</td>
<td>.47 (.36)</td>
<td>.45 (.29)</td>
<td>.38 (.28)</td>
</tr>
<tr>
<td>$M$</td>
<td>.49 (.32)</td>
<td>.45 (.27)</td>
<td>.34 (.27)</td>
</tr>
</tbody>
</table>

**4.2.5 Follow-up on Failure of the Motivation Manipulation: Was Motivation Correlated with Better Memory Performance, Independent of Condition?**

The previous analyses clearly demonstrated that the motivational manipulation did not have its intended effect on study behavior or memory performance. It is still possible, however, that highly-motivated learners performed better on the final memory
test, regardless of the condition to which they had been assigned. This possibility was investigated by calculating the bivariate Pearson product-moment correlation between self-reported motivation level and final test performance (Table 7). Indeed, there was a small but significant correlation between participants’ self-reported motivation levels and their score on the final test, $r(153) = .16, p < .05$. This result suggests that higher levels of motivation may have helped participants to perform better, independently of whether they had received the high- or low-motivational instructions. Note, however, that this interpretation is only speculative; it is impossible to determine a causal relationship on the basis of this analysis.

A few other interesting patterns are also shown in Table 7. There was not a significant correlation between self-reported motivation level and performance on the initial test, $r(153) = .06, p > .05$, nor was there a significant correlation between motivation level and time spent studying the feedback $r(153) = .15, p > .05$. However, both initial test performance [$r(153) = .70, p < .001$] and feedback study-time [$r(153) = .43, p < .001$] were correlated with final test performance. Learners who performed well on the initial test, and who spent more time studying the feedback, showed better final memory for the word pairs.
Table 7: Correlations between self-rated motivation levels and the key dependent variables in the experiment.

Note: A single asterisk indicates that the correlation is significant at the .05 level (two-tailed); a double asterisk indicates that it is significant at the 0.01 level.

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Initial test score</td>
<td></td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Motivation rating</td>
<td>.06</td>
<td></td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>3. Feedback viewing time</td>
<td>.01</td>
<td>.15</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>4. Final test score</td>
<td>.70**</td>
<td>.16*</td>
<td>.43**</td>
<td>---</td>
</tr>
</tbody>
</table>

4.2.6 Final Test Performance: Maintenance of Initially Correct Responses

One question mark raised by the current experiment is why delaying feedback did not improve later memory, as it has in dozens of previous laboratory studies (e.g., Anderson, Kulhavy, & Andre, 1972; Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962; Butler & Roediger, 2008; Carpenter & Vul, 2011; Metcalf, Kornell, & Finn, 2009; Kulhavy & Anderson, 1972; More, 1969; Phy & Andre, 1981; Rankin & Trepper, 1978; Sassenrath & Yonge, 1968, 1969; Smith & Kimball, 2010; Surber & Anderson, 1975; Sturges & Crawford, 1964). As described earlier, one prominent theory of feedback-timing effects— the spacing hypothesis— proposes that delaying feedback benefits memory because it promotes the maintenance of initially correct responses (Butler, Karpicke, & Roediger; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Smith, Kimball, & Mann, 2007). In the current experiment, initial test performance was very low, with only about one-quarter of items answered correctly.
Perhaps an analysis that focused only on those initially correct items would reveal a benefit of delaying feedback. That is, although delaying feedback did not benefit overall performance on the final test, it may have helped learners to maintain their initially correct answers.

To examine this possibility, the proportion of correct answers that were produced on the initial test and then reproduced on the final test was computed separately for each learner\(^2\). Overall, when learners produced a correct answer on the initial test, they were likely to produce it again later; however, the maintenance of initially correct responses was not perfect (\(M = .69\)). Moreover, in contrast to the predictions of the spacing hypothesis, there was no evidence that delaying feedback promoted the maintenance of initially correct answers \([F(2, 134) = 2.38 \ p = .10, \ MSE = .09, \ \eta^2 = .03 \) for the main effect of feedback-timing condition]; in fact, if anything, there was again a small hint that immediate feedback was more effective than delayed (Table 8). Motivational instructions did not affect the maintenance of initially correct answers, and there was no evidence of an interaction between motivational instructions and feedback timing (both \(Fs < 1\)).

\(^2\) Note that some participants did not produce any correct answers on the initial test; they were excluded from this analysis. Note further that excluding these same participants from all of the preceding analyses did not change any of the key outcomes of this study.
### Table 8: Proportion of initially correct answers maintained on the final test, presented separately for each of the six between-subjects conditions.

*Note: Standard deviations are in parentheses.*

<table>
<thead>
<tr>
<th></th>
<th>Immediate</th>
<th>Delayed 1</th>
<th>Delayed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low motivation</td>
<td>.76 (.24)</td>
<td>.69 (.24)</td>
<td>.67 (.32)</td>
</tr>
<tr>
<td>High motivation</td>
<td>.75 (.33)</td>
<td>.69 (.33)</td>
<td>.57 (.31)</td>
</tr>
<tr>
<td>M</td>
<td>.76 (.28)</td>
<td>.69 (.29)</td>
<td>.62 (.32)</td>
</tr>
</tbody>
</table>

#### 4.3 Discussion

Overall, the results of Experiment 5 were not in line with expectations. The biggest surprise was that the motivational manipulation did not affect memory performance. Despite rating themselves as being more motivated to learn the vocabulary words, learners who received the high-motivation instructions did not perform any better on the final test. Indeed, the motivational manipulation did not produce any discernable change in behavior; there was also no difference in study time as a function of whether learners had received the high- or low-motivation instructions.

Why did the motivational instructions fail to produce their intended effect? One possibility is that the manipulation was simply not potent enough. Indeed, participants reported relatively high levels of motivation (an average rating of close to 4, the “Fairly Motivated” point on the scale) regardless of whether they had received the high- or low-motivation instructions. A related (and not incompatible) possibility is that the effects of
the instructions—although they were strong enough to produce an initial difference in participants’ self-reported motivation ratings—had dissipated by the time participants viewed the feedback. In the immediate feedback condition, learners viewed the feedback immediately after receiving the instructions; in the delayed feedback conditions, however, the motivational instructions and the feedback viewing were separated by a delay of 10 minutes. The reasoning for placing the motivational manipulation before this delay is that we had hoped to show that the high-motivation instructions prevented a natural decline in motivation levels that would have otherwise taken place over those 10 minutes. Instead, it seems that the delay may have been enough to wipe out any effects of the manipulation.

Despite the failure of this particular motivation manipulation, however, the present experiment still provides evidence that highly-motivated learners tend to perform better on memory tasks. Participants’ self-reported motivation levels were positively correlated with their performance on the final memory test, independent of whether they had received the high- or low-motivation instructions. There was not a significant correlation between self-reported motivation levels and feedback viewing time, however. Thus, it does not seem to be the case that higher levels of motivation led participants to spend more time studying the feedback.

A different puzzle, besides the fact that the motivation manipulation did not affect study behavior or final memory performance, is why delaying feedback did not improve
memory. As previously discussed, dozens of published laboratory studies, as well as Experiments 1 and 2 of this dissertation, have found beneficial effects of delaying feedback on later memory. One possible explanation is that initial test performance was extremely low, meaning that the major function of the feedback was to correct errors rather than to reinforce already-correct responses. According to the spacing hypothesis—which thus far represents the strongest contender for explaining feedback-timing effects—delaying feedback benefits memory because it promotes the retention of initially correct answers. In the present situation, however, there were very few initially correct answers to maintain, which may explain why benefits of delaying feedback did not occur. According to this interpretation, however, a benefit of delaying feedback should have emerged in the follow-up analysis that included only those items that participants initially answered correctly. Although this result did not occur, note that the follow-up analysis was based on a very small number of items and should be taken with a grain of salt.

In retrospect, it may also be the case that delaying feedback failed to benefit memory because the schedule for the delayed feedback conditions was not planned according to the optimal spacing function characterized by Smith and Kimball (2010; see Introduction section). Specifically, the delay between the initial test and the presentation of the feedback (i.e., interstudy interval) was 10 minutes long, and learners were expected to remember the critical information for a retention interval of either 10 minutes (delayed feedback 1 condition) or 20 minutes (delayed feedback 2 condition). Thus, the length of
the interstudy interval was either 50% or 100% that of the retention interval. According to Smith and Kimball’s (2010) data, however, optimal memory performance occurs when the interstudy interval is 10-20% of the retention interval. Thus, a shorter interstudy interval (or longer retention interval) may have promoted better memory performance in the present experiment. Future researchers should take the relationship between feedback delay and retention interval into account when planning their experimental procedures.

To conclude, one interesting implication of the present results is that the level of initial performance might be important in determining whether delaying feedback helps, hurts, or hinders memory. Future studies should investigate the possibility that immediate feedback may benefit memory when initial performance is extremely low (and there are many errors to be corrected), while delaying feedback may be more beneficial when initial performance is higher (and the correct information needs to be practiced over time). It is difficult to evaluate this possibility on the basis of the existing literature, where initial test performance is typically high. If such a pattern were obtained, it would provide strong evidence for the spacing hypothesis as an explanation for feedback-timing effects.
5. General Discussion

There is no debate over the importance of providing feedback to improve student learning (e.g., Hattie, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Shute, 2008). Researchers and educators do not yet have a clear understanding, however, of how to design the most effective feedback procedures to correct specific types of errors under specific circumstances. The goal of my dissertation was to investigate the optimal procedures for administering feedback to improve memory outcomes, with a particular focus on when feedback should be administered. When considered as a package, several key findings came out of this work. First, the results provided evidence against the traditional view that delaying feedback is always harmful. Second, these experiments helped to identify several key variables that may help to determine the optimal timing of feedback under different circumstances. Third, the data provided tentative support for the spacing hypothesis (although, as noted later, this hypothesis does not seem to explain all feedback-timing effects).

As reviewed in detail in the Introduction, the claim that immediate feedback benefits learning (and is superior to detailed feedback) has a long history of support. This belief is grounded in the behaviorist learning tradition and in dozens of animal studies showing that immediate feedback is necessary to establish the contingency necessary for learning. The tide has begun to change, however, as increasing amounts of evidence demonstrate that delaying feedback often benefits human memory in laboratory studies.
By showing that benefits of delaying feedback can also be replicated in “messy”
classroom environments, Experiments 1 and 2 represent a further advance in the shifting
understanding of feedback-timing effects. Historically, researchers and educators have
seen delayed feedback as an unfortunate reality of classroom learning. This thinking is
outdated; a substantial body of evidence now suggests that, under the right circumstances,
delaying feedback can actually improve student performance.

Of course, delaying feedback does not always benefit memory; Experiments 3, 4,
and 5 of this dissertation illustrate some of the circumstances in which immediate
feedback is at least as effective. In Experiments 3 and 4, immediate feedback promoted
better correction of false memories; these high-confidence errors sometimes went
completely unnoticed when the feedback was delayed. In Experiment 5, immediate and
delayed feedback procedures were equally effective in promoting the learning of foreign
language translations. Here, it was probably not the case that the specific errors were
difficult to notice when the feedback was delayed. Instead, the sheer quantity of errors
(and the corresponding low proportion of initially correct answers) meant that the
primary function of the feedback was to correct learners’ mistakes, something that
immediate feedback seems to do at least as well as delayed feedback.

One can imagine other possible situations, not directly tested in the present
experiments, in which immediate feedback would likely outperformed delayed feedback.
For example, errors that reflect deep conceptual misunderstandings may be best corrected
immediately. Imagine a learner who is struggling to apply one or more specific rules or principles while completing massed practice on a given topic (e.g., an algebra student who, in attempting to calculate the slope of various lines, continually misapplies the slope formula). In this situation of conceptual misunderstanding, delaying feedback makes no sense; the learner will simply continue to fail while attempting the entire assignment. Instead, immediate feedback—which allows correction of the underlying conceptual error, so that the rest of the assignment can be completed correctly—is more likely to be beneficial. Note that a different learner completing the exact same activity might benefit from a different feedback timing procedure, however. For example, consider a student who progresses through the same assignment, calculating most of the slopes correctly while making only a few minor calculation errors. For this learner, delaying feedback may be beneficial; it allows spaced exposure to the slope formula that he or she already understands well.

5.1 Implications for Theoretical Accounts of Feedback Timing Effects

The spacing hypothesis provides a satisfactory account for the present data. To recap, this hypothesis assumes that delaying feedback benefits memory by allowing the learner to receive spaced exposure to items that were initially answered correctly (Butler, Karpicke, & Roediger; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Smith, Kimball, & Mann, 2007). This hypothesis is compatible with the results of Experiments 1 and 2, in which initial homework performance was relatively high and the feedback messages
provided reinforcement of those initially correct items. It is also compatible with Experiments 3, 4, and 5, which failed to demonstrate a benefit of delaying feedback when correcting false memories and learning foreign language vocabulary, respectively. Specifically, initial performance in these experiments was relatively low and so there were fewer initially correct items available to benefit from spaced practice.

Although the present results are consistent with the spacing hypothesis, however, it is unlikely that this perspective accounts for all possible data involving feedback timing. For example, some studies have shown benefits of even very small delays of feedback, on the order of a few seconds (e.g., Brackbill & Kappy, 1962; Carpenter & Vul, 2011; Mullaney, Carpenter, Grotenhuis, & Burianek, 2014; Schroth, 1992; Sturges, Sarafino, & Donaldson, 1968). In these situations, it seems unlikely that delaying feedback provides a meaningful amount of spaced practice in comparison to an immediate feedback condition. Indeed, rather than appealing to the spacing hypothesis to explain these results, Carpenter and Vul (2011) and Mullaney et al. (2014) proposed a curiosity-based explanation. Specifically, they hypothesized that delaying the feedback by just a few seconds causes the learner to experience a sense of anticipation, or curiosity to learn the correct answers. Thus, they are more likely to pay attention when the correct answer is presented. Consistent with this hypothesis, Carpenter and Vul (2011) showed that the benefit of briefly delaying feedback disappeared when learners were required to complete an unrelated counting task during the delay interval. They hypothesized that the
unrelated task distracted learners from engaging in anticipatory processing, thus wiping out the effects of delay. Note that the spacing hypothesis does not easily explain the same pattern; there seems no reason why a filled delay would eliminate potential benefits of spaced practice.

5.2 Practical Implications

In some ways, the nuanced answer to the question “When should feedback be administered?” could evoke a feeling of resignation. Indeed, even if educators were made aware of all of the possible complex factors (e.g., the level of initial performance, the type of error to be corrected, the likelihood of viewing the feedback, etc.) that likely contributing to determining the most effective feedback timing procedure, it would still not be practically feasible for them to deliver an optimal feedback message at an optimal time to each of their individual students on a or concept-by-concept basis.

Considering the current boom in educational technology, however, achieving such a personalized approach to learning becomes much more realistic. Intelligent tutor systems, e-books, and online learning applications can all be designed with optimal feedback delivery procedures in mind. Importantly, these procedures can be tailored to the individual learner; students can receive immediate feedback on some concepts and delayed feedback on others, depending on factors including previous performance and expected retention interval. With this possibility in mind, achieving a thorough
understanding of how to use feedback to benefit human and learning memory is a more worthy goal than ever.
## Appendix A

Responses to survey questions in Experiment 1 and 2.

### Experiment 1

<table>
<thead>
<tr>
<th>Question</th>
<th>Timing of Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate (N = 11)</td>
</tr>
<tr>
<td>How did you feel about receiving feedback [immediately / after a delay]?</td>
<td></td>
</tr>
<tr>
<td>I liked it a lot</td>
<td>82%</td>
</tr>
<tr>
<td>I liked it a little</td>
<td>18%</td>
</tr>
<tr>
<td>I did not like it</td>
<td>0%</td>
</tr>
<tr>
<td>In terms of your performance in the course, do you feel like you benefited from [immediate / delayed] feedback?</td>
<td></td>
</tr>
<tr>
<td>I benefited a lot</td>
<td>73%</td>
</tr>
<tr>
<td>I benefited a little</td>
<td>27%</td>
</tr>
<tr>
<td>I did not benefit</td>
<td>0%</td>
</tr>
<tr>
<td>It hurt my performance</td>
<td>0%</td>
</tr>
<tr>
<td>When you received feedback, how did you use it to help you learn?</td>
<td></td>
</tr>
<tr>
<td>I did not look at it</td>
<td>0%</td>
</tr>
<tr>
<td>I looked at how many problems I got correct / incorrect</td>
<td>0%</td>
</tr>
<tr>
<td>I looked at the solution for problems that I got incorrect</td>
<td>64%</td>
</tr>
<tr>
<td>I looked at the solution for all problems</td>
<td>36%</td>
</tr>
</tbody>
</table>
### Experiment 2

<table>
<thead>
<tr>
<th>Question</th>
<th>Required (N = 17)</th>
<th>Optional (N = 12)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Did you prefer receiving feedback immediately or after a delay?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I highly preferred immediate feedback</td>
<td>94%</td>
<td>83%</td>
</tr>
<tr>
<td>I slightly preferred immediate feedback</td>
<td>0%</td>
<td>8%</td>
</tr>
<tr>
<td>I did not have a preference</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>I slightly preferred delayed feedback</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>I highly preferred delayed feedback</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>In terms of your performance in the course, do you feel like you</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>benefited more from either immediate feedback or delayed feedback?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I benefited a lot more from immediate feedback</td>
<td>76%</td>
<td>83%</td>
</tr>
<tr>
<td>I benefited a little more from immediate feedback</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>I benefited equally from immediate and delayed feedback</td>
<td>18%</td>
<td>8%</td>
</tr>
<tr>
<td>I benefited a little more delayed feedback</td>
<td>0%</td>
<td>8%</td>
</tr>
<tr>
<td>I benefited a lot more from delayed feedback</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>On weeks when you received immediate feedback, how did you</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>use the feedback to help you learn?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I did not look at it</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>I looked at how many problems I got correct / incorrect</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>I looked at the solution for problems that I got incorrect</td>
<td>29%</td>
<td>50%</td>
</tr>
<tr>
<td>I looked at the solution for all problems</td>
<td>59%</td>
<td>42%</td>
</tr>
<tr>
<td><strong>On weeks when you received delayed feedback, how did you use</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the feedback to help you learn?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I did not look at it</td>
<td>18%</td>
<td>33%</td>
</tr>
<tr>
<td>I looked at how many problems I got correct / incorrect</td>
<td>12%</td>
<td>17%</td>
</tr>
<tr>
<td>I looked at the solution for problems that I got incorrect</td>
<td>24%</td>
<td>33%</td>
</tr>
<tr>
<td>I looked at the solution for all problems</td>
<td>47%</td>
<td>17%</td>
</tr>
</tbody>
</table>
References


Little, J. K. (1934). Results of use of machines for testing and for drill, upon learning in educational psychology. *Journal of Experimental Education, 3*, 45-49.


Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*, 41-44.
Biography

Born

December 19th, 1988 in Roanoke, VA

Education

Duke University, Durham, NC
Ph.D., Psychology & Neuroscience, 2016

Duke University, Durham, NC
M.A., Psychology, 2014

Furman University, Greenville, SC
B.A., Majors in Psychology and Philosophy, 2011

Publications


**Honors and Awards**

<table>
<thead>
<tr>
<th>Year</th>
<th>Award Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Psychonomics Graduate Student Travel Award for Outstanding Research</td>
</tr>
<tr>
<td>2014</td>
<td>Curran-Bauer Analytics Scholarship for Advanced Statistics Training</td>
</tr>
<tr>
<td>2012-2015</td>
<td>National Science Foundation Graduate Research Fellow</td>
</tr>
<tr>
<td>2011</td>
<td>The Outstanding Senior Psychology Major</td>
</tr>
<tr>
<td>2007-2011</td>
<td>Mickel Presidential Scholar</td>
</tr>
<tr>
<td>2007-2011</td>
<td>National Merit Scholar</td>
</tr>
</tbody>
</table>