SPACING AND THE DELAY-RETENTION EFFECT: AN ALTERNATIVE EXPLANATION OF THE EFFECTS OF FEEDBACK TIMING ON SEMANTIC LEARNING

by

TROY ANTHONY SMITH

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN PSYCHOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2007

ACKNOWLEDGEMENTS

I would like to thank my faculty mentor, Dr. Daniel Kimball, for helping me learn to think like an experimental psychologist. His training in critical thinking and theoretical analysis, experimental design, and statistical analysis has proved invaluable; without his training, I would not have known how to even begin this project. I would also like to thank him for asking tough questions and providing help whenever I needed it at every stage of this project, from inception to execution to write-up. His tough, insightful questions encouraged me to think deeply and critically about the relevant theories and about my own arguments. His comments and feedback have immeasurably improved the quality of my experiments and have helped hone my scientific writing skills.

Other people who deserve thanks for their role in the completion of this thesis include the other members of my committee, Dr. Timothy Odegard and Dr. Ira Bernstein, and the other graduate students who have been part of Dr. Kimball's lab, Tony Coker and Nancy Bonner. Tony has been a great sounding board for ideas, and his comments and questions have inspired and encouraged me throughout this project. Nancy worked with me extensively on a previous experiment that laid the foundation for the ideas and experiments reported in this thesis; the central idea for this thesis was sparked in one of our many conversations about that project. Finally, were it not for the help of the undergraduate research assistants who worked in Dr. Kimball's lab during the fall of 2006 and spring of 2007, I would probably still be collecting data the data for Experiment 1!

July 12, 2007

ABSTRACT

SPACING AND THE DELAY-RETENTION EFFECT:
AN ALTERNATIVE EXPLANATION OF THE
EFFECTS OF FEEDBACK TIMING
ON SEMANTIC LEARNING

Publication No. ______

TROY ANTHONY SMITH, M.S.
The University of Texas at Arlington, 2007

Supervising Professor: Daniel R. Kimball, J.D., Ph.D.

Current theoretical accounts of feedback timing effects on retention are problematic. Some predict that delayed feedback should lead to better retention; others predict that immediate feedback should lead to better retention. Previous empirical findings are unclear: Some studies have found an advantage for delayed feedback, some an advantage for immediate feedback, and some no difference. In three experiments involving new semantic learning, I tested the extent to which spacing and lag effects can account for these seemingly contradictory findings, based on predictions of the new theory of disuse (Bjork & Bjork, 1992). Experiment 1 compared the effects of timing variations for repeated study trials, repeated test trials, and feedback trials. Experiments 2 and 3 examined the effects of restudy and retest trials following immediate and delayed feedback, and the impact of varying study-feedback lag. Results support the spacing hypothesis and challenge competing theories of feedback timing.

TABLE OF CONTENTS

Appendix

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Technological advances are revolutionizing education by providing efficient methods to tailor instruction to the needs of individual students, decreasing the barriers to delivery of that instruction, and creating new and more efficient tools for tracking student performance (Mayer, 2005). But these advances are also raising important theoretical questions about the nature of learning and practical questions about how to best apply theoretical learning principles so as to maximize the efficiency of the learning process and the long-term retention of information. One critically important aspect of learning theory that needs to be better understood is the effect of feedback on semantic learning. In a recent review, Mory (2004) challenged researchers to critically re-examine the effects of feedback in light of new pedagogical perspectives and new theoretical understandings of learning.

An ideal perspective from which to re-examine the effects of feedback on semantic learning is the "new theory of disuse" developed by R. A. Bjork and his colleagues (Bjork & Bjork, 1992; Bjork & Bjork, 2006; Schmidt & Bjork, 1992). The new theory of disuse is a functional description of how human memory operates, and the underlying theoretical constructs have been validated by numerous empirical studies (for a review, see Bjork & Bjork, 2006). Additionally, the new theory of disuse provides a set of relatively simple principles to guide the transition from theory to practice. This mixture of theoretical assumptions that have been empirically validated and principles for practical application derived from the theoretical assumptions gives the new theory of disuse the ability to explain certain observed learning phenomena that cannot be satisfactorily explained by other theories of learning as well as the ability to generate unique predictions for laboratory and applied research.

This paper focuses on the so-called “delay-retention effect”—the finding that delaying feedback results in an improvement in performance on tests after a long retention interval but not necessarily after a short retention interval—and why this effect is observed in some cases but reversed in other cases. The paper also briefly addresses the interplay between respondents’ confidence in their answers on a test and the effect of feedback following the test. Chapter 2 reviews the key empirical findings regarding the effect of variations in the timing of feedback on later test performance along with the theoretical explanations that have been offered to explain the key findings. Chapter 3 summarizes the assumptions of the new theory of disuse and offers possible alternative explanations of the previously reviewed empirical findings based on these assumptions. Chapter 4 briefly reviews the major studies that have examined how confidence on a test impacts the effect of feedback and discusses the servocontrol theory of feedback that has been used to explain these effects. Finally,Chapter 5 describes a series of three experiments that were designed to test the predictions of the new theory of disuse and the previously reviewed theories.

CHAPTER 2

IMMEDIATE VS. DELAYED FEEDBACK: A REVIEW

2.1 Operational Definitions

The fact that standard operational definitions for feedback timing have never been established and the fact that researchers have often failed to distinguish between different types of memory tests are major sources of confusion in the feedback timing literature. For example, some researchers who have defined feedback provided a few minutes after a test as immediate feedback (e.g., Kulhavy & Anderson, 1972; Phye & Andre, 1989; Webb, Stock, & McCarthy, 1994) have generally found that delayed feedback leads to better performance. However, other researchers who have defined the exact same manipulation as delayed feedback (e.g., Beeson, 1973; Buzhardt & Semb, 2002; Clariana, 2000) have failed to find any advantage for delaying feedback and in some cases have found an advantage for immediate feedback. Clearly, the use of different operational definitions by the researchers complicates attempts to compare these studies to each other. Therefore, before reviewing the literature, it is important to distinguish between the different operational definitions and testing procedures that have been used by various researchers over the last 60 years.

2.1.1 Feedback Timing

In general, an experimenter can focus on item-level events (such as study trial, test trial, feedback trial) or on experimental-level events (such as study phase, test phase, feedback phase). The choice of operational definitions for any particular experiment will be highly dependent on the event scale of interest to the experimenter. For example, a researcher who is interested in testing specific theoretical predictions would likely focus on item-level events and thus might operationalize immediate feedback as feedback presented after the test trial for each item. On the other hand, a researcher interested in classroom applications would likely focus on experiment-level events and thus might operationalize immediate feedback as feedback presented after the test phase for all items. This difference in focus would naturally lead to different operational definitions, which would then in turn make it difficult to compare these studies. One obvious way to reduce the resulting confusion is to include the level of focus in the operational definition. For clarity, I will use the following operational definitions for feedback timing throughout the remainder of this paper.

2.1.1.1 Item-by-item immediate (IBI-I) feedback

The cue\question and correct response are provided to the participant after each test item with no intervening events such as study or test trials for the same item or for other items. Note that an unfilled delay on the order of microseconds or a few seconds (e.g., Rankin & Trepper, 1978, Group D) may be introduced between the participant's response and presentation of the correct response.

2.1.1.2 Item-by-item delayed (IBI-D) feedback

The cue\question and correct response are provided to the participant for each test item after intervening events such as study or test trials for other items. The length of the filled delay may be based on elapsed time or on number of intervening items. In studies that use IBI-D feedback, the amount of elapsed time between the participant's response and presentation of the correct response is usually on the order of several minutes (e.g., Rankin & Trepper, 1978, Group LD).

2.1.1.3 End-of-test immediate (EOT-I) feedback

The cue\question and correct response to all of the test items are provided to the participant immediately after the participant has completed the entire test. Optimally, there should be no elapsed time between completion of the test and presentation of feedback, and there should be no intervening events. In practice, the elapsed time is often on the order of 1-30 minutes, during which the participants may or may not be given any specific task (e.g., Newman, Williams, & Hiller, 1974).

2.1.1.4 End-of-test delayed (EOT-D) feedback

The cue\question and correct response to all of the test items are provided to the participant after the participant has completed the entire test and a specific amount of time has elapsed. The elapsed time should be substantially longer than the elapsed time in the contrasting EOT-I condition. In practice, the EOT-D feedback delay is often 24-48 hours (e.g., Kulhavy & Anderson, 1972, TDF condition).

2.1.2 Testing Procedures

Most theories of memory distinguish between recognition processes that are tested by forced-choice tests such as those using multiple-choice questions and recall processes that are tested by cued recall and free recall tests (Gillund & Shiffrin, 1984; Mandler, 1980; Tulving, 1983). However, most feedback timing studies have failed to make any theoretical or operational distinction between recognition and recall. Feedback timing effects on long-term retention have been studied almost exclusively using multiple-choice tests (for reviews, see Kulik & Kulik, 1988; Mory, 2004). In fact, only two published studies of which I am aware have used a cued recall test to examine the effect of variations in feedback timing on retention (see Table 2.1). Unless otherwise noted, all of the studies discussed in the following review used multiple-choice tests. The potential problems and confounds that can be introduced by multiple-choice tests are addressed later in this section.

Table 2.1. Feedback Timing Studies with Retention Intervals of at Least 1 Day

		Course	Test	Feedback Timing[3]		Feedback	Retention	
Study[1]	Setting	Content?	Type[2]	Immediate	Delayed	delay	interval[4]	Results
Angell (1949)								
	Classroom	Yes	MC	IBI-I + IBI-D	IBI-D	3-4 days	weeks	Immediate
Beck & Lindsey (1979)								
	Classroom	Yes	MC	EOT-I	EOT-D	7 days	weeks	Null
Beeson (1973)								
	Classroom	Yes	MC	IBI-I	EOT-D	unknown	weeks	Immediate
Brackbill & Kappy (1962)								
	Lab	No	2AFC	IBI-I	IBI-I	5s, 10s	1 day	Delayed
							8 days	Null
Brackbill, Bravos, & Starr (1962)								
Study 1	Lab	No	2AFC	IBI-I	IBI-I	5s, 10s	1 day	Delayed
							8 days	Delayed
Study 2	Lab	No	2AFC	IBI-I	IBI-I	10s	1 day	Delayed
							8 days	Delayed
Brosvic, Dihoff, Epstein, & Cook (2006)								
Study 1	Classroom	Yes	MC	IBI-I	EOT-I	EOT	days	Immediate
Study 2	Classroom	Yes	MC	IBI-I	EOT-I	EOT	days	Immediate
Brosvic, Epstein, Cook, & Dihoff (2005)								
	Classroom	Yes	MC	IBI-I	EOT-I	0 hrs	2-10 weeks	Immediate
					EOT-D	24 hrs		
Buzhardt & Semb (2002)								
	Teaching lab	Yes	MC+TF	IBI-I	IBI-D	EOT	weeks	Null
Clariana, Wagner, & Murphy (2000)								
	Lab	No	MC	IBI-I	IBI-D	EOT	24 hrs	Null

Continued on Next Page...

Table 2.1 - continued

		Course	Test	Feedback Timing		Feedback	Retention	
Study[1]	Setting	Content?	Type[2]	Immediate	Delayed	delay	interval[4]	Results
Dihoff, Brosvic, & Epstein (2003)								
	Classroom	Yes	MC	IBI-I	EOT-I	0 hrs	2-14 weeks	Immediate
					EOT-D	24 hrs		
Dihoff, Brosvic, Epstein, & Cook (2004)								
Study 2	Classroom	Yes	MC	IBI-I	EOT-I	0 hrs	2-12 weeks	Immediate
					EOT-D	24 hrs		
Dihoff, Brosvic, Epstein, & Cook (2005)								
Study 1	Classroom	Yes	MC	IBI-I	EOT-I		1-5 days	Immediate
English & Kinzer (1966)[5]								
	Unknown	Unknown	MC	IBI-I	EOT-D	7 days	Unknown	Delayed
Haynes (1974)								
	Classroom	No	MC	EOT-I	EOT-D	24 hrs	7 days (I) 6 days (D)	Null
Kulhavy & Anderson (1972)								
Study 1	Lab	No	MC	EOT-I	EOT-D	24 hrs	7 days (I) 6 days (D)	Delayed
More (1969)								
	Classroom	Yes	MC	IBI-I	EOT-D	(2.5hrs,	0 days	Delayed
						1 day,	3 days	Delayed
						4 days)		
Newman, Williams, & Hiller (1974)								
	Classroom	Yes	MC	IBI-D (≈25 min)	IBI-D	(1 day, 7 days)	7 days	Null
Paige (1966)								
	Classroom	Yes	PS	EOT-I + EOT-D	EOT-D	24 hrs	3 weeks	Immediate
Phye & Andre (1989)								
	Lab	Yes	MC	EOT-I	EOT-D	24 hrs	24 hrs	Delayed[6]
Phye & Baller (1970								
	Lab	Yes	MC	IBI-D	IBI-D	48 hrs	0 days	Null
				(≈30min)			7 days	Null

Continued on Next Page...

Table 2.1 - continued

		Course	Test	Feedback Timing		Feedback	Retention	
Study[1]	Setting	Content?	Type[2]	Immediate	Delayed	delay	interval[4]	Results
Phye, Gugliemella, & Sola (1976)								
Study 1	Lab	No	MC	EOT-I	EOT-D	48 hrs	0 days	Null
							7 days	Null
Study 2	Lab	No	MC	EOT-I	EOT-D	48 hrs	0 days	Null
							7 days	Null
Pressey (1950)								
Study 1	Classroom	Yes	MC	IBI-I	EOT-D	24 hrs	weeks	Null
Study 2	Classroom	Yes	MC	IBI-I	EOT-D	24 hrs	weeks	Immediate
Rankin & Trepper (1978)								
	Lab	No	MC	IBI-I	IBI-I	15s	24 hrs	Delayed
					IBI-D	5-10 min		
Robin (1978)[5]								
	Unknown	Unknown	Unknown	IBI-I	EOT-D	48 hrs	Unknown	Immediate
Sassenrath & Yonge (1968)								
	Classroom	Yes	MC	IBI-D	IBI-D	24 hrs	0 days	Null
				(≈15 min)			5 days	Delayed
Sassenrath & Yonge (1969)								
	Classroom	Yes	MC	IBI-I	IBI-I	10s	0s	Null
							5 days	Delayed
Sturges (1969)								
	Lab	No	MC	IBI-I	IBI-D	24 hrs	0 days	Null
							7 days	Delayed[6]
Sturges (1972)								
Study 1	Lab	No	MC	IBI-I	IBI-D	(20 min,	0 days	Delayed
						24 hrs)	7 days	Delayed
Study 2	Lab	No	MC	IBI-I	IBI-D	(20 min,	0 days	Null
						24 hrs)	7 days	Delayed[6]
Sturges (1978)								
	Teaching lab	Yes	CR+MC	IBI-I	IBI-D	(20 min, 24 hrs)	1-3 weeks	Delayed[6]
Sullivan, Schultz, & Baker (1971)								
	Classroom	Yes	MC	IBI-I	EOT-D	2-5 days	2-16 days	Immediate

Continued on Next Page...

Table 2.1 - continued

		Course	Test	Feedback Timing		Feedback	Retention	
Study[1]	Setting	Content?	Type[2]	Immediate	Delayed	delay	interval[4]	Results
Surber & Anderson (1975)								
	Classroom	No	MC	EOT-I	EOT-D	24 hrs	7 days (I) 6 days (D)	Delayed
Swindell & Walls (1993)								
	Lab	No	CR	IBI-D (≈10 min)	IBI-D	24 hrs	8 days (I) 7 days (D)	Delayed[6]
Webb, Stock, & McCarthy (1994)								
Study 1	Lab	No	MC	IBI-D (≈30 min)	IBI-D	24 hrs	7 days (I) 6 days (D)	Delayed[6]
Study 2	Lab	No	MC	IBI-D (≈30 min)	IBI-D	24 hrs	7 days (I) 6 days (D)	Null

[1] Each of the studies in this table examined the relative effects of immediate and delayed feedback on retention of semantic learning and were published in peer-reviewed journals.

[2] CR = constructed response (i.e., cued recall)

MC = multiple choice

PS = problem solving

TF = true-false

2AFC = two alternative forced-choice discrimination

[3] EOT-D = end-of-test delayed feedback

EOT-I = end-of-test immediate feedback

IBI-D = item-by-item delayed feedback

IBI-I = item-by-item immediate feedback

See Section 2.1.1 for specific definitions.

[4] Some studies used different retention intervals for the immediate (I) and delayed (D) feedback conditions.

[5] Based on data reported by Kulik & Kulik (1988)

[6] Reported results are questionable due to null effects for some analyses.

2.2 Early Views on the Effect of Feedback Timing

Some of the earliest studies of the effects of feedback timing in semantic learning were conducted at a time when behaviorism and operant conditioning formed the basis for learning theory. Classical learning theories based on stimulus-response-reinforcement stipulate that the function of feedback is to reinforce correct responses and that the effectiveness of feedback in doing this should be optimal when it is provided in close temporal proximity to the response (Hull, 1952; Saltzman, 1951; Skinner, 1938, 1957). According to this theoretical perspective, delaying feedback should lead to a decrease in performance and retention compared to providing immediate feedback (Pressey, 1950; Renner, 1964). Numerous studies with animals and humans in simple stimulus-response studies have demonstrated that delaying feedback during learning trials does indeed lead to poorer performance on later test trials than is observed when immediate feedback is given (Renner). However, the vast majority of these studies involved learning behaviors rather than learning semantic knowledge. Nevertheless, based on conditioning theory, behaviorist principles, and these comparative studies, it was implicitly assumed that the same results would apply to semantic learning (Renner). Indeed, the entire programmed instruction movement and other attempts to automate and improved learning such as Pressey's "mechanical teacher" were based on this assumption (Pressey, 1963), despite the fact that very few studies had specifically examined feedback timing effects in semantic learning.

Angell (1949) conducted one of the few studies of the time that did attempt to test this assumption. Angell divided the students enrolled in an undergraduate chemistry course into an immediate feedback group and a delayed feedback group. Both groups were given 3 quizzes throughout the semester and received the appropriate type of feedback on each quiz. Students in the immediate feedback group used punch-boards that revealed the correct answer to each question on an item-by-item basis (IBI-I) after they had recorded their answer for that question, while students in the delayed feedback group took the same quizzes but did not receive any feedback during the quizzes. Instead, during the next class meeting, the instructor went over the exam question-by-question, thus providing feedback to the delayed feedback (IBI-D) group—and providing an additional study opportunity to the immediate feedback group. At the end of the semester, all of the students took a final exam that covered the same course material as had been covered by the quizzes but did not use the same questions. Perhaps not surprisingly, the immediate feedback group performed significantly better on the final exam than did the delayed feedback group. Angell interpreted this finding as support for the superiority of immediate feedback in promoting learning, as predicted by classical conditioning theory. Of course, the finding could be due to the fact that the immediate feedback group had an extra spaced restudy opportunity compared to the delayed feedback group.

By the 1960's, some researchers were beginning to question whether the principles of learning stimulus-response behaviors could realistically be applied to semantic learning (Pressey, 1963; Renner, 1964). As Pressey, one of the pioneers of the programmed instruction movement, declared:

> The whole trend of American research and theory as regards learning has been based on a false premise–that the important features of human learning are to be found in animals. Instead, the all-important fact is that human has transcended animal learning. Language, number, such skills as silent reading, make possible facilitations of learning, and kinds of learning, impossible even for the apes. (pg. 5)

2.3 The Delay-Retention Effect

2.3.1 Effect of Short Feedback Delays

In 1962, Brackbill and her colleagues (Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962) published the results of a series of experiments that indicated that contrary to predictions from operant condition theory, delaying feedback can actually facilitate retention, at least in some circumstances. In these experiments, elementary school children were presented with pairs of pictures (Brackbill, Bravos, et al.; Brackbill & Kappy) or pairs of meaningless letter bigrams (Brackbill, Issacs, et al.) in a series of two alternative forced-choice discrimination learning trials. On each learning trial, feedback was presented following the item with a delay of 0s, 5s, or 10s[1]. In the initial experiment (Brackbill & Kappy), delaying feedback by a few seconds resulted in better performance over a 1 day retention interval, but no difference was found between any of the conditions over an 8 day retention interval. In the follow-up experiments, delaying feedback by a few seconds facilitated performance over both retention intervals. Brackbill, Bravos, et al. termed their finding the *delay-retention effect.*

Sassenrath and Yonge (1969) extended the delay-retention effect results to semantic learning by using multiple-choice exams with undergraduate students. Because they were using course content material, they did not include a study phase; instead, the participants took a timed multiple-choice test over material that had been previously covered in class. Feedback was provided immediately after each question or after an unfilled delay of 10s[2]. All participants were given an immediate multiple-choice retest

[1] All three of these manipulations would be considered IBI-I feedback by the operational definitions presented earlier.

[2] Both of these manipulations would be considered IBI-I feedback by the operational definitions presented earlier.

and a delayed multiple-choice retest 5 days later. Sassenrath and Yonge found that delaying feedback by a few seconds did not have any significant effect on the immediate retest but did lead to better retention after the relatively long 5-day retention interval.

These findings that waiting to provide feedback for a few seconds after a participant responds to each item leads to an improvement in retention was theoretically interesting because it was a clear violation of the temporal contiguity principles of contemporary theories of learning (e.g., Hull, 1952; Skinner, 1957). Sassenrath and Yonge (1969) speculated that temporal contiguity principles might not apply to semantic learning because language gives humans the unique ability to associate temporally disparate events. Brackbill, Bravos, et al. (1962) hypothesized that the delay allowed both the overt and subsequent covert responses that would occur during the delay to be reinforced whereas immediate feedback only reinforced the overt response (see also, Renner 1964). A similar explanation consistent with more recent theories of memory (e.g., Hintzman, 1984; Raaijmakers & Shiffrin, 1981; Tulving, 1983) is that the unfilled delay allows for additional processing time which participants can use to continue rehearsing the items for which they are reasonably certain they responded correctly. This additional rehearsal time would cause items in the delayed feedback condition to have stronger traces in memory, on average, than items in the immediate feedback condition. The difference in trace strengths would then lead to the the corresponding difference in performance on a retention test. From this perspective, the initial delay-retention effect findings (Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962; Sassenrath & Yonge, 1969) are unsurprising and perhaps even trivial. Also, these initial findings are of minimal practical value to educators, with the application being limited to revising the timing in programmed instruction lessons (O'Day, Kulhavy, Anderson, & Malczynski, 1971).

2.3.2 Effect of Longer Delays

However, subsequent findings that have shown a delay-retention effect with feedback delays on the order of 24-48 hours are much more interesting. These findings are interesting to theoreticians because they raise questions that are challenging for classic and modern theories of learning and memory. They also have potentially important and widespread implications for education. Educators can easily introduce a 24-48 hour feedback delay in a classroom environment by returning exams and quizzes to students on the next class meeting and going through the exam question-by-question; in fact, many educators already use this technique on a routine basis.

One of the first studies to find a delay-retention effect with a longer feedback delay was conducted by Sassenrath and Yonge (1968). They compared the effect of providing feedback after participants completed an exam (EOT-I feedback) to the effect of delaying feedback for 24 hours (EOT-D feedback) and found that delaying feedback by 24 hours did not have a significant effect on performance for an immediate retest but did lead to better performance on a test after a 5-day retention interval. These results must be interpreted with caution however, as the difference between mean performance of the immediate and delayed feedback groups on the delayed retest was relatively small in absolute terms (3.5 percentage points, Cohen's $d = 0.53$) and performance on the tests was near ceiling (92% and 94% correct for immediate EOT-I and delayed EOT-D feedback, respectively, on the immediate retests; 85% and 88% correct for the corresponding delayed retests).

Kulhavy and Anderson (1972) replicated this basic finding in a well-controlled study that did not have the potential problems with ceiling effects. Kulhavy and Anderson divided a large group of high school students who had not taken any psychology classes into eight groups in which the sequence and timing of events were varied. These groups included two immediate feedback (EOT-I) groups and two 24-hour delayed feedback (EOT-D) groups. On the first day of the experiment, all of the students in these groups took an initial multiple-choice test consisting of questions from a college-level introductory psychology course. They then received feedback in the form of the questions and the correct answers at the designated time—either immediately after completing the test or the next day—and were allowed to restudy the questions and answers for up to an hour. One week after taking the initial test, the students were given a final delayed test; unlike in the previously discussed studies, the students did not take an intervening immediate test. Kulhavy and Anderson found that students in the delayed feedback (EOT-D, 24 hrs) group correctly answered significantly more questions (M = 66%) on the final delayed retention test that did students in the immediate feedback (EOT-I) group (M = 50%). Several other studies have also found similar results using multiple-choice tests with immediate end-of-test feedback compared to delayed end-of-test feedback (e.g., Phye & Andre, 1989; Webb, Stock, & McCarthy, 1994, Experiment 1).

The finding that longer, filled delays between test and feedback also leads to a delay-retention effect raises an important theoretical question: Can the effect of a long delay be explained by the same mechanisms as the effect of a short delay? The stimulus-response reinforcement explanation offered by Brackbill, Bravos, et al. (1962) does not seem to be a sufficient explanation for the effect of a 24-hour delay. While it is reasonable to suppose that a short delay on the order of a few seconds might enhance stimulus-response reinforcement, it seems highly unlikely that the much longer 24-hour delay would result in a similar enhancement. Indeed, it seems much more likely that the extremely long delay would impair reinforcement. Neither does the differential rehearsal explanation seem to be sufficient to explain the effect of the longer, filled delay; an effective filled delay practically eliminates the possibility of differential rehearsal[3].

[3] Sassenrath and Yonge (1968) speculated that the participants in the delayed feedback condition used the time between the initial test and the feedback to rehearse the correct responses and that this rehearsal somehow interacted with the retention interval to produce improved performance on the

2.4 The Interference-Perseveration Hypothesis

Kulhavy and Anderson (1972) introduced the first viable theoretical explanation for the delay-retention effect with long feedback delays—the interference perseveration hypothesis. They noted that the pattern of results in the typical delayed feedback experiment is similar to the pattern observed in interference experiments using the AB-AC[4] paradigm (for a recent review, see Anderson & Neely, 1996; for a contemporaneous review, see Postman & Underwood, 1973). In these experiments, participants start by studying a list of cue-target paired associates (the AB list). They then study a second list in which the original cues are re-paired with new targets (the AC list). After a specified retention interval, participants are given a cued recall test in which they are either instructed to recall the targets from a specific list, the target that first came to mind (a modified free recall, MFR, task) or both targets (a modified modified free recall, MMFR, task). The typical finding in AB-AC experiments is that for short retention intervals recall of the C responses is generally superior to recall of the B responses but for longer retention intervals recall of the B responses increases relative to recall of the C responses. Over time, the earlier learned responses appear to interfere with the ability to recall the later learned responses (Postman & Underwood).

Kulhavy and Anderson (1972) claimed that this proactive interference is responsible for the delay-retention effect. In their words:

> Our explanation is very simple: learners forget their incorrect responses over the delay interval, and thus there is less interference with learning the correct

delayed retention test but not on the immediate retention test, but they did not provide any possible theoretical mechanisms that might give rise to this hypothesized interaction.

[4] The AB-AC paradigm is generally referred to as the AB-AD paradigm in the more recent interference literature such as Anderson and Neely (1996). I am using the older term to maintain consistency with the feedback timing literature.

> answers from the feedback. The subjects who receive immediate feedback, on the other hand, suffer from proactive interference because of the incorrect response to which they have committed themselves. This explanation will be called the interference-perseveration hypothesis. (pg. 506)

The interference perseveration hypothesis is intellectually attractive because at first glance it appears to provide a simple, parsimonious account for the delay-retention effect on semantic learning that is consistent with well established principles of cognitive psychology. This intellectual attractiveness has enabled the interference perseveration hypothesis to become the leading theoretical account of the delay-retention effect for over 30 years (Mory, 2004).

However, a closer examination of literature reveals that there are severe methodological, theoretical, and empirical problems that call into question not only the interference perseveration hypothesis as a theoretical explanation of the delay-retention effect, but also the robustness and generality of the effect itself. Some methodological and theoretical problems stem from the near exclusive use of multiple choice tests in previous studies. Questions have also been raised with regard to the evidence for perseveration of initial errors—the central mechanism of the interference perseveration hypothesis. Finally, many studies of feedback timing with semantic material have failed to find evidence for the delay-retention effect, raising doubt as to the robustness and generality of the effect. Each of these difficulties are discussed in detail below.

2.4.1 Use of Multiple-Choice Tests

With only two exceptions (Sturges, 1978; Swindell & Walls, 1993), every peer-reviewed study of feedback timing effects on the retention of semantic information of which I am aware has used multiple choice tests for the initial test and for the final assessment (see Table 2.1). Sturges used both constructed response (i.e., short answer or cued recall) questions and multiple choice questions but failed to find any effect of feedback timing for the constructed response questions. Swindell and Walls measured participants' memory for information from a text passage using constructed response tests with unelaborated feedback (correct answer only), elaborated feedback (correct answer with an explanation), or self-generated feedback (re-presentation of the original text passage). They found a significant delay-retention effect for the self-generated feedback condition but failed to find significant significant effects of feedback timing for the other conditions. Thus, every study that has found a reliable effect of feedback timing with feedback delays longer than a few seconds has used multiple-choice tests. This nearly exclusive use of multiple-choice tests is problematic for two reasons.

First, the use of multiple-choice tests places important methodological limitations on mapping the AB-AC paradigm to the delayed feedback paradigm. Multiple-choice tests are a type of forced-choice recognition test, but the AB-AC paradigm traditionally involves cued recall tests (Anderson & Neely, 1996). While proactive interference has been shown to occur for recognition memory using 2 alternative forced-choice discrimination tests (Dean, Garabedian, & Yekovich, 1983; Kane & Anderson, 1978; Neill & Mathis, 1998; Petrusic & Dillon, 1972), the size of proactive interference effect with these type of recognition tests is relatively small compared to the effect on recall (Anderson & Neely). Further, no studies of proactive interference with recognition memory have been performed in which the distractors were present during the initial learning trials, as is the case in the standard delayed feedback experiments using multiple-choice tests.

Second, when multiple-choice tests are used during the learning phase, they can introduce errors in initial encoding of information, they can interfere with previously learned information, and they can enhance memory for the selected responses (Roediger & Marsh, 2005; Roediger & Karpicke, 2006a, 2006b). This makes it particularly important to be aware the procedures that were used in any particular feedback timing study when interpreting the results of the study, because using a different event sequence can dramatically impact whether errors are encoded prior to the multiple-choice test or during the test. If a study-test-feedback-test sequence is used, then the multiple-choice test itself can act as a source of retroactive interference. In this case, determining whether any feedback timing effects that may be observed are due to testing, proactive interference, retroactive interference, or some complex interaction between these effects can be quite challenging, if not impossible. If there is no initial study trial (i.e, a test-feedback-test sequence is used, as was done by Kulhavy & Anderson, 1972) then the results may be more theoretically interpretable; however, this type of procedure does not seem very ecologically valid if one is interested in applying the results to an educational setting.

2.4.2 Error Perserveration

One of the largest problems with the interference perseveration hypothesis is that there is scant evidence for the central mechanism that is posited to be responsible for the delay-retention effect. In order to say that the delay leads to more forgetting of initial errors thus making them more likely to be corrected, one must examine what happens to initial errors in each of the feedback timing conditions by tracking each possible type of response at each test stage, including initial correct and error responses (R_1 and W_1, respectively), repetition of an initial error on the second test (W_{old}), new errors that are made on the second test (W_{new}), and correct responses on the second test (R_2). After the responses are categorized,a conditional analysis can be conducted to determine the probability of an initial error being corrected, $P(R_2|W_1)$, the probability of an initial error being repeated, $P(W_{old}|W_1)$, and the probability of a new error being made, $P(W_{new}|W_1)$. Distinguishing between old errors being repeated (error perseveration) and new errors being made is a critical, but often neglected, portion of the conditional analysis. For example, the conditional analyses conducted by Kulhavy and Anderson (1972) and by

Surber and Anderson (1975)—key studies that have been used to support the interference perseveration hypothesis—did not discriminate between new errors and old errors. Thus, their measure of error perseveration, the simple probability $P(W_2|W_1)$, was faulty because it counted new errors the same as old errors. Studies that have performed a complete conditional analysis wherein old and new errors were separated from each other have found no evidence that the original errors from the initial test perseverate to the final test any more in the immediate feedback condition than in the delayed feedback condition (Brosvic et al., 2005; Clariana et al., 2000; Peeck, 1979; Phye and Andre, 1989) as the interference perseveration hypothesis predicts.

2.4.3 Inconsistent Results

Another major problem with the interference perserveration hypothesis is that it cannot account for the large number of studies that have failed to find any difference in performance between immediate and delayed feedback conditions, even with retention intervals on the order of several months, or for the studies that have found an advantage for immediate feedback (see Table 2.1). One might explain the null results as the effect of unintentional release of proactive interference (Kincaid & Wickens, 1970) due to irregularities in the experimental procedures used, but an examination of the procedures used in many of these studies does not reveal any systematic factor that might allow for such a release of proactive interference. The studies used a variety of materials such as course content, vocabulary word-definition pairs, arithmetic facts, and trivia facts with adults and with children. The experimental manipulations included using item-by-item immediate feedback, end-of-test immediate feedback, and feedback delays varying from 24 hours to 7 days. The only things in common among these studies was the use of multiple-choice tests and the comparison of the relative effects of immediate and delayed feedback on performance on a later test of retention.

Explaining the findings that immediate feedback leads to better performance is somewhat easier, because most if not all of these studies have allowed for differential rehearsal opportunities, either by design (e.g., Angell, 1949) or by loose experimental control (e.g., Dihoff et al., 2003). When participants in the immediate feedback conditions are afforded study opportunities that are not afforded to the participants in the delayed feedback conditions, it should not be surprising that the immediate feedback group performs better on a retention test.

2.5 Other Theoretical Perspectives

Although the interference perseveration hypothesis is currently the leading theoretical account of the effect of feedback timing on semantic learning, it has not gone unchallenged. Clariana et al. (2000) have proposed a modern discrepancy reduction theory using a connectionist model based on a delta rule of learning. The model predicted that immediate feedback should lead to better long-term retention than would delayed feedback. However, this prediction did not agree with the results that Clariana et al. obtained in an empirical test of the model. Epstein and his colleagues (Brosvic, Dihoff et al., 2006; Brosvic, Epstein et al., 2006; Dihoff et al., 2003; Dihoff et al., 2004) have sharply criticized the interference perseveration hypothesis on the basis that it ignores the error correction function of feedback. They have presented evidence that providing immediate item-by-item feedback in classroom settings can lead to higher rates of error correction than delaying feedback until the end of test or until the next day, which in turn leads to higher scores on a later retention test. However, these results must be interpreted with caution because the experimental designs did not control for restudy opportunities; participants in the immediate feedback conditions could have restudied test material during the feedback delay, despite Epstein and his colleagues' attempts to minimize differential study opportunities. Also, Epstein and his colleagues appear to be more interested in promoting the efficacy of providing immediate feedback to students taking a test in a classroom setting—particularly when done so using a proprietary tool they have developed (Epstein & Brosvic, 2002a, 2002b; Epstein, Epstein & Brosvic, 2001; Epstein et al., 2002)—than they are in understanding the theoretical aspects of feedback timing. While clearly demonstrating that the interference perseveration hypothesis is an insufficient theoretical account of feedback timing, they barely sketch the outlines of possible alternative theoretical accounts.

2.6 Summary

In contrast to predictions from early theories of learning based on stimulus-response reinforcement, Brackbill and her colleagues (Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962) found that delaying feedback for a few seconds could increase retention compared to providing feedback immediately after each test item. Subsequent studies such as those by Sassenrath and Yonge (1968) extended the finding to feedback delays on the order of 24-48 hours. Kulhavy and Anderson (1972) offered the interference-perseveration hypothesis to explain these findings, and it has been the leading theoretical account of feedback timing effects ever since. However, there are multiple problems with the interference-perseveration hypothesis. Chief among these problems is the fact that the predicted delay-retention effect has been found in fewer than one-third of feedback timing studies. Alternative accounts of the findings in the literature such as discrepancy reduction theories or the differential attention hypothesis leave much to be desired.

In short, the empirical findings regarding the effect of variations in the timing of feedback on semantic learning are anything but clear, and the theoretical accounts that have been offered to date cannot explain the wide variety of findings in the literature. Of course, explaining the diverse array of findings would be challenging for any theory.

Perhaps, the simple, elegant explanation offered by Kulhavy and Anderson (1972) is too simple. It may be that a more general theory is needed. In the next chapter, I outline the basic assumptions of a general theory of learning and memory that may fit the bill quite well—the new theory of disuse (Bjork & Bjork, 1992)—and describe how it can be applied to feedback timing studies.

CHAPTER 3

THE NEW THEORY OF DISUSE

The new theory of disuse was formulated in response to the following three observed "peculiarities" of human memory: (a) the storage capacity of memory appears to be nearly unlimited while the retrieval process exhibits severe limitations; (b) retrieval actually modifies memory as much, if not more than, initial encoding; (c) and the accessibility of items in memory regresses over time (Bjork & Bjork, 1992). Using these empirical observations as a starting point, Bjork and Bjork laid out a set of five theoretical assumptions which taken together can explain a large number of empirical phenomena, including the above noted peculiarities of human memory. In this section, I summarize these assumptions, discuss the predictions of the new theory of disuse with regard to learning and forgetting, and extend the application of the theory to the feedback timing paradigm.

3.1 Assumptions of the New Theory of Disuse

Before summarizing the assumptions of the new theory of disuse, it is important to note that Bjork and Bjork (1992) focus on describing the functional operation of memory without grounding the new theory of disuse in any particular view of the architecture of memory. Indeed, they often describe the new theory of disuse as a "framework" (pg. 58) rather than as a model of memory. Although the definitions and assumptions of the new theory of disuse are described using quantitative terminology, they have not yet been specified and implemented in a quantitative model of memory. It remains to be seen whether these functional descriptions can be implemented in a quantitatively specified memory model. Nevertheless, the assumptions provide a powerful means of explaining a number of memory phenomena—including regression and recovery of learned information and behaviors, effects of overlearning and repeated learning, and effects of distributed practice (Bjork & Bjork, 1992, 2006)—and may be useful in understanding effects of feedback timing variations.

3.1.1 Assumption 1: Distinct Storage and Retrieval Strengths

Consistent with several classical accounts of learning (e.g., Skinner, 1938; Hull, 1943; Estes, 1955), the new theory of disuse assumes that the representation of information in memory can be indexed by two interdependent strengths. *Storage strength* is an index of how well learned a given representation is; it has been described as "a latent variable that has no direct effect on performance" (Bjork & Bjork, 1992, pg. 42). *Retrieval strength*, on the other hand, is an index of how accessible an item in memory is when given a specific set of retrieval cues and has a direct effect on performance. Storage strength and retrieval strength are assumed to be distinct from each other but interact as described in Assumptions 2, 4 and 5 (Bjork & Bjork).

If memory is viewed as an associative network (e.g., Collins & Loftus, 1975), then storage strength can be thought of as an index of the number and strength of connections between the node representing a particular concept and its associated nodes, and retrieval strength can be thought of as the activation of a particular node in response to a given cue set. Note that at any given point in time, the storage strength of a concept has a single fixed value, but the retrieval strength of a concept will take on different values depending on the cue set that is used to search memory. Because retrieval strength for a given item is governed by the relationship between the item and the retrieval cues that are used, the probability of recalling a particular item is assumed to be completely determined by the retrieval strengths of its representations in relation to other representations that are associated with the same set of cues (see also Assumption 3).

3.1.2 Assumption 2: Storage Strength Increases Monotonically

Storage strength is assumed to be a monotonically increasing function that grows with each opportunity to study an item and with each successful recall of an item. The storage strength function theoretically can increase without bounds, and, as a consequence, the storage capacity of long-term memory is virtually unlimited. However, the increment in storage strength due to additional restudy or recall attempts is limited by two factors. First, the increment decreases as the current storage strength increases, causing the storage strength function to be negatively accelerated. Second, the increment in storage strength for a given item in memory is inversely related to the retrieval strength of that item at the time of a restudy event or a successful retrieval during a test event (Bjork & Bjork, 1992). From an associative view of memory, the increment in storage strength could be a reflection of adding new associative links due to initial encoding or to elaboration; it could be a reflection of strengthening existing associative links; or it could be a combination of both of these processes.

3.1.3 Assumption 3: Limitations of Retrieval Capacity

The new theory of disuse assumes that there are two components in the recall process that place limits on retrieval strength: discrimination and reconstruction. The discrimination component directly implements retrieval competition, and the reconstruction component implements forgetting effects (Bjork & Bjork, 1992). These components are conceptually similar to the sampling and recovery rules respectively in the Search of Associative Memory model (Raaijmakers & Shiffrin, 1981).

When memory is probed with a given set of cues, a number of different representations in memory will be activated, and these representations will compete with each other for recall. In order for a particular item to be recalled, its associated representation(s) must be discriminated from the competing representations activated by the cue set. This is done using a standard ratio rule: The probability of discriminating a specific representation can be expressed as the ratio of its retrieval strength compared to the sum of the retrieval strengths of all representations activated by the cue set. Assuming that the representation for a specific item was successfully discriminated, the associated item must then be reconstructed from the representation. The probability of reconstruction for a specific item is a function of the absolute retrieval strength of that item. Thus, a representation that is highly discriminable because it has a high retrieval strength relative to other representations that were activated by the current cue set might nonetheless fail to be recalled if it also has a low absolute retrieval strength (Bjork & Bjork, 1992).

So, in contrast to storage capacity—which is assumed to be unlimited—retrieval capacity is limited by retrieval competition. The consequence of this limitation is that adding new items to memory or increasing the retrieval strength of some items can decrease the relative retrieval strength of other items, thus making them less likely to be recalled. In particular, adding a new item that is associated to the same retrieval cues as an existing item makes the existing item less retrievable because it reduces the relative discriminability of the existing item. Similarly, when two items are associated to the same set of retrieval cues, increasing the retrieval strength of one of those items decreases the relative discriminability of the other item (Bjork & Bjork, 1992).

3.1.4 Assumption 4: Increments in Retrieval Strength

Restudy and test events give rise to increments in retrieval strength, with the increment due to a successful retrieval being greater than the increment due to restudy.

The retrieval strength increment is an increasing function of the current storage strength and a decreasing function of the current retrieval strength for the retrieved item. The higher the current storage strength of a successfully retrieved item is, the larger the retrieval strength increment; conversely, the lower the current retrieval strength is, the larger the increment (Bjork & Bjork, 1992).

3.1.5 Assumption 5: Decrements in Retrieval Strength

Learning or retrieval of other items that are associated to the same cues as a given item is assumed to result in a decrement of the retrieval strength of that item. The decrement function for retrieval strength is a mirror of the increment function. The lower the current storage strength of a successfully retrieved item is, the larger the retrieval strength decrement. Conversely, the higher the current retrieval strength is, the larger the decrement (Bjork & Bjork, 1992).

3.1.6 Illustration of the Assumptions

Figure 3.1 illustrates the assumptions of the new theory of disuse, with changes being represented as vectors. Horizontal vectors represent storage strength increments, and vertical vectors represent changes in retrieval strength. According to Assumption 2, the increment in storage strength is a decreasing function of current storage strength (horizontal vectors b and d are smaller than horizontal vectors a and c, respectively) and the current retrieval strength (horizontal vectors a and b are smaller than horizontal vectors c and d, respectively). According to Assumption 4, the increment in retrieval strength is an increasing function of the current storage strength (vertical vectors b and d are larger than vertical vectors a and c, respectively) and a decreasing function of the current retrieval strength (vertical vectors a and b are smaller than vertical vectors c and d, respectively). Finally, according to Assumption 5, the decay in retrieval strength

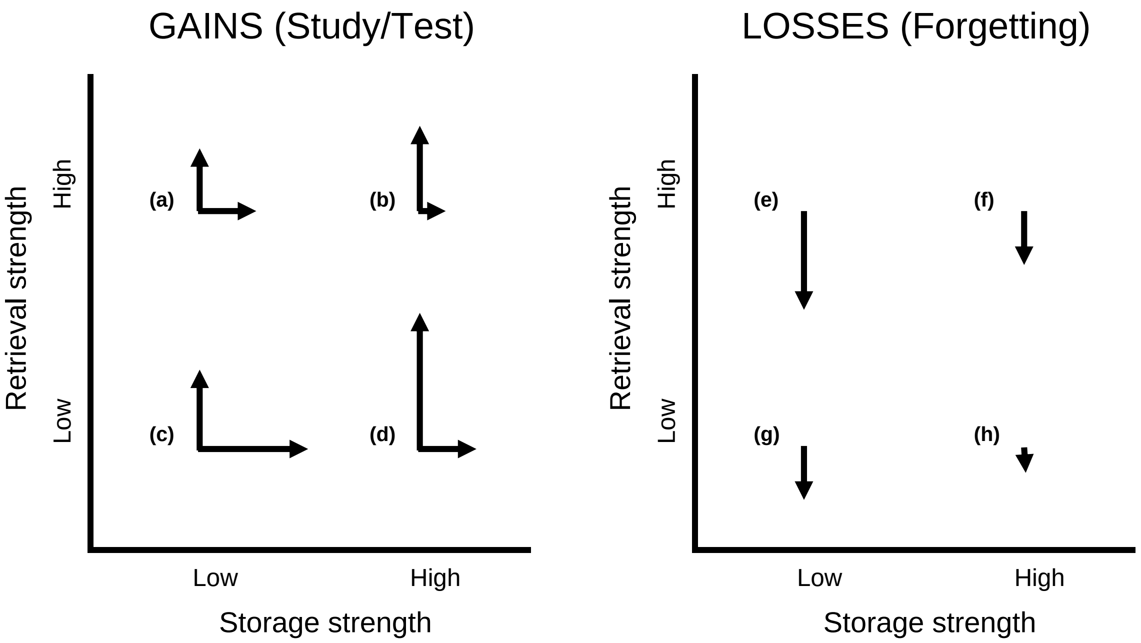

Figure 3.1. Graphical summary of the rules for changes to storage and retrieval strengths in the new theory of disuse. Horizontal vectors represent storage strength increments. Vertical vectors represent changes in retrieval strength.

(forgetting) is a decreasing function of the current storage strength (f and h are smaller than e and g, respectively) and an increasing function of the current retrieval strength (e and f are larger than g and h, respectively). Figure 3.1 also shows the interaction of retrieval strength and storage strength. For example, vector a shows that studying an item that has low storage strength but high retrieval strength—such as the name of a person you just met—results in only moderate increments to the storage and retrieval strengths; however, studying an item that has high storage strength and low retrieval strength—such as the name of someone you knew well 20 years ago—results in a large boost in retrieval strength, as can be seen in vector d.

3.2 Predictions of the New Theory of Disuse

3.2.1 Forgetting and Learning Curves

The new theory of disuse easily accounts for the hypothetical forgetting curve (Figure 3.2). Because absolute retrieval strength is assumed to decay over time and with intervening events (Assumption 5), the probability of discrimination and reconstruction—and therefore the probability of retrieval—also decays. If a representation is not activated by continuing study or test events, the retrieval strength will eventually decay to the point where the representation cannot be retrieved (i.e., the item can be considered to be completely forgotten). It is important to note, however, that according to the new theory of disuse, forgetting is not caused by decay of the information in memory—after all, the storage strength has not decreased—but rather by the fact that the item is not currently retrievable because it cannot be discriminated from other representations activated by the same cue or because it no longer has sufficient absolute retrieval strength to be reconstructed (Bjork & Bjork, 1992).

The new theory of disuse can also account for the hypothetical learning curve for repeated study trials. The representation of a concept is activated in memory each time the concept is studied, and the storage and retrieval strengths are incremented on each successful study trial (Assumption 4). As the retrieval strength increases, the item becomes easier to retrieve, the retrieval strength increment decreases, and the recall probability approaches one (Assumption 4). With a constant inter-trial interval, the new theory of disuse predicts the typical learning curve shown in Figure 3.3. Overlearning effects are also predicted due the the fact that while retrieval strength is bounded, storage strength is not. Additional study beyond a perfect performance criteria increases storage strength without increasing retrieval strength; this increase in storage strength slows the

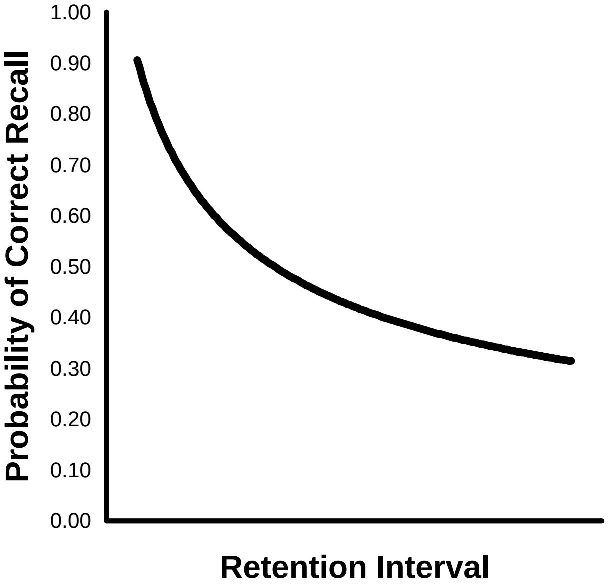

Figure 3.2. Hypothetical forgetting curve.

decay of retrieval strength (Assumption 5) and, hence, the observed rate of forgetting (Bjork & Bjork, 1992).

3.2.2 Desirable Difficulties

Because the storage strength increment is a decreasing function of current retrieval strength, the new theory of disuse predicts that manipulations that result in lower retrieval strengths on learning trials will lead to better retention than those that result in higher retrieval strengths, as long as the retrieval strengths in both manipulations are sufficiently high to allow the correct representation to be retrieved from memory on a large proportion of learning trials. Bjork (1994) has referred to these retrieval-strength lowering conditions as *desirable difficulties.* From the perspective of the educator or trainer, they are are desirable because they lead to improvements in long-term retention. However, these conditions also make the learning process more difficult for the student or trainee because they make it harder to retrieve the appropriate responses from memory

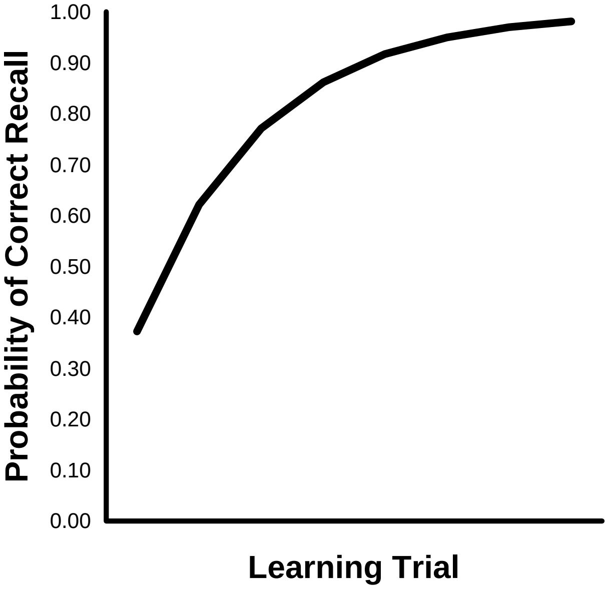

Figure 3.3. Hypothetical learning curve.

during learning trials. Examples of desirable difficulties include spacing or distributing practice, reducing feedback to the learner (at least in motor learning tasks), using tests as learning events, using an expanding retrieval practice schedule, varying the conditions of practice, and introducing contextual interference (Bjork, 1994; Bjork & Bjork, 2006). Because distributed practice effects play an important role in the experiments presented later in this paper, a detailed account of the predictions the new theory of disuse makes with regard to spacing and lag effects is in order.

3.2.2.1 The Spacing Effect

The term *spacing effect* refers to the finding that spaced study episodes lead to better learning than do massed study episodes (for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1989, 1996; Janiszewski, Noel, & Swayer, 2003). The spacing effect has been studied extensively since the time of Ebbinghaus (1885) and is highly robust. For example, 259 of the 271 comparisons of spaced versus massed study in verbal memory tasks examined by Cepeda et al. found a significant performance advantage for spaced presentations.[1] The benefit for spaced study has been consistently observed across a wide range of retention intervals, from a few seconds[2] to as long as eight years (Bahrick & Phelps, 1987); with a wide array of stimuli, including nonsense syllables, word lists, paired-associates, pictures, and text (Janiszewski, et al.); and in almost every verbal learning paradigm, including recognition, cued recall, and free recall (Dempster, 1996).

The new theory of disuse easily predicts that, in general, distributed practice (study or test trials) will lead to better retention than will massed practice (Bjork & Bjork, 1992, 2006). With spaced trials, forgetting occurs between practice trials (Assumptions 3 and 5), whereas little or no forgetting occurs between massed trials. Thus, the retrieval strength on a given spaced trial will be lower than the retrieval strength on the corresponding massed trial. The difference in retrieval strengths, in turn, causes the storage strength increment due to a successful retrieval on a spaced trial to be larger than the storage strength increment on the corresponding massed trial (Assumption 2; also, compare vectors c and a in Figure 3.1). The resulting higher storage strengths in the spaced condition retards forgetting (Assumption 5; compare vectors h and g in Figure 3.1) which leads to better long-term retention.[3]

[1] Each of the studies included in Cepeda et al.'s (2006) analysis equated total study time between the spaced and massed conditions, thus ruling out differential study time as an explanation for the findings.

[2] Cepeda et al. (2006) have noted that the "Peterson paradox"—the finding that massed study leads to better performance than does spaced study for short retention intervals (Peterson, Hillner, & Saltzman, 1962; Peterson, Saltzman, Hillner, & Land, 1962)—appears to be much more constrained than has previously been believed, with the effect being limited to spacing intervals of 4-8s and retention intervals no greater than 8s.

[3] Over sufficiently short retention intervals in which there is little or no forgetting between spaced trials, the new theory of disuse predicts that massed study may lead to better performance than spaced

3.2.2.2 Lag Effects

The term *lag effect* refers to changes in performance for different levels of spacing. Lag effects are not nearly as well understood as the basic spacing effect, and until recently many studies have not distinguished between the spacing effect and lag effects (Cepeda et al., 2006). Unlike the spacing effect, lag effects are highly sensitive to changes in retention interval, with the spacing interval (lag) and the retention interval interacting in rather complex ways. The meta-analysis conducted by Cepeda et al. showed that lag effects are generally nonmontonic (Figure 3.4). As long the the spacing interval is not long relative to the the retention interval, increases in the spacing interval lead to increases in performance on a final retention test (the left-hand portion of the curve in Figure 3.4). For any given retention interval, there is a spacing interval that maximizes performance (the peak of the curve in Figure 3.4), and increases in spacing interval beyond this optimal value result in decreases in performance (the right-hand portion of the curve in Figure 3.4). Cepeda et al. also found that the optimal spacing interval increases as the retention interval increases.

The new theory of disuse provides a simple explanation for the shape of the lag effect curve: Increasing the interval between distributed practice sessions increases the forgetting that occurs between sessions (i.e., decreases the retrieval strength). As long as items in memory remain retrievable, this lower retrieval strength translates to a larger increment in storage strength. The larger storage strength, in turn, results in a slower rate of forgetting, which manifests as an improvement in performance on a retention test. An optimal spacing interval is one that results in most of the items barely being retrievable on each practice trial. However, if the interval between practice trials is too

study because massed study can lead to higher initial retrieval strengths (Bjork & Bjork, 2006). This may explain the "Peterson paradox" (Peterson, Hillner, & Saltzman, 1962).

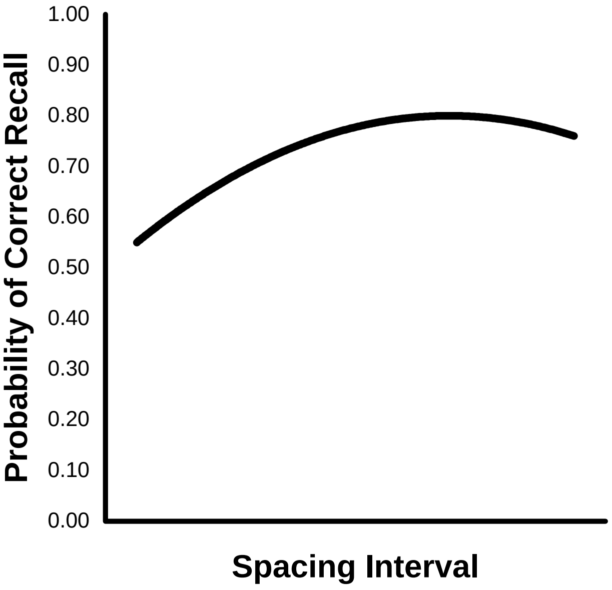

Figure 3.4. Hypothetical lag effect curve.

long, the retrieval strength becomes so low that many items cannot be retrieved on the practice trials. These non-retrieved items do not receive boosts in storage or retrieval strength, and their retrieval strengths continue to decay, making them even less likely to be retrieved in the future. Thus, as the spacing interval increases beyond the optimal value, retention test performance should decline.

3.3 Applying the New Theory of Disuse to Feedback Timing

The new theory of disuse has not previously been applied to the effects of feedback[4], but at this point in the discussion, this is a straightforward exercise. By definition, delayed feedback is a spaced study event, and, as detailed above, the new theory of dis-

[4] Bjork (1994) has noted that providing summary feedback after several trials rather than providing feedback on every trial in motor skill learning can be a desirable difficulty based on empirical studies. However, Bjork did not provide any theoretical explanation of feedback effects based on the new theory of disuse, nor did he address the issue of feedback timing with verbal learning.

use predicts that, in general, spacing study opportunities improves long-term retention. Therefore, the new theory of disuse predicts that, in general, delayed feedback will improve long-term retention. In other words, the delay-retention effect is a distributed practice effect and can be explained by the new theory of disuse in exactly the same way that spacing and lag effects with repeated study trials are explained. I refer to this idea as the *spacing hypothesis* throughout the remainder of this paper. Of course, there are two obvious aspects of the feedback timing paradigm that differ from the typical spaced study paradigm and may modulate spacing and lag effects: testing effects and error correction effects.

3.3.1 Testing Effects

The term *testing effect* refers to the finding that taking a test over previously learned material improves performance on a later test more so than does simply restudying the material (Dempster, 1996). The testing effect is a fairly robust phenomena and has been observed in a wide variety of laboratory and field studies (for a review, see Roediger & Karpicke, 2006a). The prediction that test events, in general, are more powerful learning opportunities than are restudy events flows naturally from Assumption 4 in the new theory of disuse. The detailed cue set during a restudy event leads to a relatively high retrieval strength compared to the more impoverished cue set during a test event. Thus a test event that ends with a successful retrieval leads to a larger increment of retrieval strength than does a simple restudy event and, consequently, increases the future retrievability of the recalled item. On the other hand, items whose retrieval strength is so weak that the item cannot be successfully retrieved with a test event cue set (i.e. items that have been "forgotten") will benefit more from a restudy event than from a retest event (Bjork & Bjork, 1992).

In addition to the basic testing effect, a *test-spacing effect* wherein spaced tests more effectively facilitate learning than do massed tests has also been observed in a number of studies (e.g., Carpenter & DeLosh, 2005; Landauer & Bjork, 1978; Whitten & Bjork, 1977; see Dempster, 1996 and Roediger & Karpicke, 2006a for reviews), and increasing the intervals between test trials has been shown to lead to a positive testing effect that is similar to the lag effects described earlier (Modigliani, 1976; Roediger & Karpicke, 2006b; Spitzer, 1939). Retrieval hypothesis accounts such as the new theory of disuse readily explain the test-spacing effect in exactly the same way as they explain the basic spacing effect: Introducing a delay between test trials makes the retrieval of information more effortful, which leads to larger increases in strength and thus to better retention (Dempster, 1996). The new theory of disuse also explains the positive testing effect as another example of a lag effect.[5]

3.3.2 Error Correction

According to the new theory of disuse, errors can only be corrected by feedback if the feedback increases the retrieval strength of the correct response enough that it can be discriminated from other possible responses on the next test opportunity. Thus, the efficacy of error correction due to feedback will be a complex function of the current storage and retrieval strengths of the item that was erroneously retrieved—that is, the strengths at the time the feedback is given—and the current storage and retrieval strengths of the correct item. Until the new theory of disuse is specified quantitatively, the exact nature of the function cannot be known and precise predictions cannot be made. Nevertheless,

[5] This implies that the lag effect curves for distributed retrieval practice will have the same shape as lag effect curves for distributed study (Figure 3.4). In other words, the positive testing effect should decline and eventually reverse with forgetting. This basic pattern can be seen in the data from Spitzer (1939).

we can make general predictions by examining the major classes of errors. For simplicity, I will divide errors into two classes: well-learned and poorly-learned.

Well-learned errors are items that have strong storage strengths but are nonetheless incorrect, such as the “fact” that opponents of Christopher Columbus thought the world was flat (Stern, 2004). In most cases, well-learned errors also have high absolute and relative retrieval strength and the discrepancy between the strengths of the correct response and the strengths of the error response are quite large. This large discrepancy implies that the correct response will often still have a lower retrieval strength than a well-learned error, even after the correct response has been incremented from studying the feedback. Thus, the probability of correcting a well-learned error is low. However, a well-learned error can still be corrected if the correct response has high enough storage and retrieval strengths prior to feedback. In this case, the feedback can increase the retrieval strength of the correct response enough to correct the error, at least on a test at a short retention interval. But even when a well-learned error is corrected, the error might reappear on a later test, in the same way that AB responses can recover over time in the AB-AC paradigm (Bjork & Bjork, 1992). This occurs when the retrieval strength of the correct response decays more than the retrieval strength of the error response.

Poorly-learned errors are items that have weak storage strengths but whose retrieval strength at test was high relative to the correct response. Most errors that are committed during new learning fall into this category because the storage strength and the relative retrieval strength of the correct response are low. Poorly-learned errors are much easier to correct than well-learned errors because even a small boost in absolute retrieval strength combined with a boost in storage strength and the consequent reduction in the decay of retrieval strength can increase the relative retrieval strength of the correct response enough so that the correct response can be discriminated from potential errors on the next test trial.

As for the effect of feedback timing on error correction, the basic prediction of the new theory of disuse is essentially consistent with the interference perseveration hypothesis: Delaying feedback increases the amount of forgetting (i.e., decreases the retrieval strength) of the incorrect response, thus making it easier for the error to be corrected. However, unlike the interference perseveration hypothesis, the new theory of disuse has specific mechanisms to make more nuanced predictions regarding the relationship between how well-learned the error is and how effective feedback will be in correcting the error. Well-learned errors are unlikely to be corrected, whether feedback is given immediately or at a delay, because the high storage strength of the error response results in a slow forgetting rate between test and feedback. Thus, the retrieval strength of the error response will change very little in that interval, and the probability of correcting the error will be essentially the same. On the other hand, poorly-learned errors, including guesses, are more likely to be corrected after delayed feedback than after immediate feedback because the low storage strength of the error response does not slow the forgetting between test and feedback. Thus, the error response will have a lower retrieval strength after a delay, which increases the probability of the error being corrected.

3.4 Explanation of Null Effects and Immediate Feedback Effects

In Chapter 2 we saw that the wide variety of results in the feedback timing literature substantially challenged the interference perseveration hypothesis and discrepancy reduction theories of feedback timing. An obvious question is whether the new theory of disuse can explain the pattern of results in a reasonable manner. As described above, the new theory of disuse predicts that, in general, delayed feedback should lead to better retention than immediate feedback. However, there are important limitations on this prediction that provide reasonable explanations for null effects and immediate feedback effects.

First, the initial study-test lag has an important impact on feedback because it affects the amount of forgetting that has occured by the time of the initial test and the effective retrieval strength during that test. If the initial study-test lags are not equivalent in the immediate and delayed feedback conditions, then the feedback timing effect will be confounded with test-spacing effects. Although this has generally not been a problem in previous studies, it is an important theoretical and practical issue. Even when there are no differences in study-test lag, it can still impact the effectiveness of feedback by altering the proportion of initial errors. With the right combination of study-test lag and test-feedback lag, the relative effectiveness of immediate and delayed feedback can be reversed. For example, suppose the study-test lag is long enough that most of the items cannot be retrieved with a test cue but can be retrieved with a study cue. Further suppose the combination of study-test lag and test-feedback lag in the delayed feedback condition is so long that many unpracticed items are completely forgotten by the time delayed feedback is given (that is, their retrieval strength is extremely low even when given highly specific cues). In this case, the items that are not answered correctly on the initial test will benefit from a spacing increment if feedback is given immediately but will not benefit from a spacing increment in the delayed feedback condition. Thus, the immediate feedback condition should lead to better retention. This may have been what happened in the study by Sullivan et al. (1971).

Second, operational definitions matter a great deal. In the new theory of disuse, IBI-I feedback is very much like massed study trial and is the only truly immediate feedback. Any delay of “immediate” feedback beyond a few seconds after the initial test response changes the nature of the effect from a spacing effect to a lag effect. Thus, a comparison of EOT-I feedback with EOT-D feedback is really a comparison of a short delay to a longer delay. In these kinds of studies, the interaction between lag and retention interval must be kept in mind. Differences in lag relative to the retention interval can cause the experimental conditions to be on opposite sides of the lag curve. Thus, in extreme cases, lag effects in and of themselves could be responsible for a reversal in the predicted effects, especially if the"immediate" feedback condition is near the peak of the lag curve while the delayed feedback condition is on the far right side of the curve (see Figure 3.4). Of course, if both feedback conditions are near the same location on the curve, then it is possible that there will be little, if any benefit to the delayed feedback condition.

Third, the new theory of disuse functionally describes the effect of differential rehearsal. As was mentioned in Chapter 2, many of the the studies that have found null or immediate feedback effects have either explicitly or implicitly given participants in the immediate feedback conditions additional study opportunities that were not given to the delayed feedback participants. These additional study opportunities cause the immediate feedback condition items to have higher storage strengths than the delayed feedback condition items, thus slowing the rate of forgetting and enhancing retention.

Finally, the new theory of disuse predicts that the current strength of information in memory impacts the effectiveness of feedback. Well-learned correct information gets a smaller boost from feedback than does poorly-learned information. Also, poorly-learned errors are easier to correct than are well-learned errors, and are more susceptible to feedback timing effects. Unless a learner makes an extra effort to correct them, well-learned errors are unlikely to be corrected, regardless of when feedback is given.

CHAPTER 4

CONFIDENCE JUDGMENTS

The effect of participants' confidence in their responses is another interesting aspect of learning with feedback that has received relatively little attention in the literature. Only one theoretical account of note has been offered to explain the effect of response confidence on the effectiveness of feedback. Further, only a handful of studies have directly examined this effect or the possible interactions between response confidence and the effect of variations in feedback timing. This chapter briefly reviews the theory and associated studies.

4.1 The Servocontrol Model of Response Certitude

The only theoretical account to specifically address the role of response confidence in the effectiveness of feedback in semantic learning is the servocontrol model of response certitude (Kulhavy & Stock, 1989). The theory starts with the assumption that the feedback paradigm taps into a closed-loop system whose main functions are detecting, analyzing, and correcting errors. This system is complex and can change its characteristic responses depending on the type, form, and content of the information that is input at each stage of the learning cycle. Unlike simple error detection schemes that rely solely on external verification to detect possible errors, the servocontrol model assumes that errors are detected by an internal comparison process that is part of the feedback cycle. In the servocontrol model, the feedback cycle consists of three interlinked phases or cycles: (I) the instructional task demand or initial test, (II) feedback processing, and (III) the criterion task demand or final test.

In Cycle I, the learner starts with the goal of generating an initial confidence judgment. This is done by comparing the perceived task demand to the individual's previous experiences with the current task or other tasks of a similar nature. This comparison also generates a set of possible responses. The learner cycles through these possible responses, comparing each one to internal and external referents. (Internal referents consist of information in episodic and semantic memory that is related to the possible responses; these can be used on any type of task. External referents are only available in recognition tasks; on a multiple-choice test, the external referents are the set of response alternatives.) As a result of the comparison, each response possibility is assigned a correctness value based on its match to the relevant referents and on the individual's initial confidence from the task demand comparison that started the cycle. The comparison process continues until the set of possible responses is exhausted or the allotted time is used up. When the comparison process is complete, the response with the highest correctness value is selected, and the response is stored in memory along with a terminal confidence judgment based on the final calculated correctness value of the selected response. This stored initial response (R1) and its associated confidence judgment become the internal referent for Cycle II.

Cycle II starts when the learner receives external feedback in the form of a correct answer. This answer is compared to the previously stored response and confidence judgment from Cycle I in order to calculate a discrepancy value. The discrepancy measures both the perceived correctness of the initial response and the confidence in that response. Positive discrepancy values represent perceived errors, and negative discrepancy values represent perceived correct responses. The absolute value of the discrepancy represents the confidence in the initial response, with higher numbers indicating higher confidence levels[1]. Thus, the learner assigns highly positive discrepancy values to responses that were thought to be correct but actually were incorrect, highly negative values to correct answers that were thought to be correct, and mildly positive or negative values to responses that the learner was unsure of and that turned out to be incorrect or correct, respectively. The servocontrol model assumes that the learner uses this discrepancy value to determine how much time and effort to expend studying the feedback message, with feedback study time and effort being directly related to the discrepancy value. While studying the feedback, the learner stores the correct response (R2) in memory along with a new confidence judgment. The strength of the new confidence judgment is assumed to be a simple function of the discrepancy value.

Cycle III starts when the original task demand is re-presented to the learner (i.e., during the final criterion test). Cycle III operates like Cycle I: The learner compares the perceived task demand to other task demands in memory, generates a set of possible responses, cycles through the responses assigning a correctness value to each, and then selects the response with the highest correctness value. As in Cycle I, correctness values are assigned to each possible response based on the degree of match between the response and available referents. The internal referent in Cycle III reflects the sum total of experience with the task: It contains the information from the Cycle I internal referent along with the responses and confidence judgments from Cycles I and II. Thus, the final response (R3) is determined in part by the relative relations of the previous confidence judgments.

An additional critical assumption is that the "durability" of a stored response is a positive function of the confidence judgment associated with that response. All other

[1] The discrepancy value can be operationalized by asking participants to make a confidence judgment (using a Likert-type scale) during the initial test (Cycle I) and then multiplying this confidence judgment by +1 if the initial test response was incorrect or by -1 if the initial test response was correct.

things being equal, responses with higher confidence judgments will be more likely to be available for future use. Curiously, Kulhavy and Stock (1989) do not discuss exactly how some items become more available and other items become less available. Unlike other theories of learning, their servocontrol model does not specify whether storage strength is variable or fixed or whether there is any decay of information in memory.

The servocontrol model makes several predictions that can be empirically tested. First, the theory predicts that the amount of time spent studying feedback will be a direct function of the discrepancy value, with larger discrepancies leading to longer feedback-processing times. For items that were correctly answered, the theory predicts that the feedback merely serves to maintain the correct response in memory. Consequently, learners are expected to spend a minimal amount of time studying high-confidence correct responses. It is not clear whether the servocontrol model allows for differential strengthening due to additional restudy, as most other theories of learning do. Second, the model predicts that the likelihood of correctly responding to an item on the final criterion test is a V-shaped function of the response confidence for that item on the initial test. Learners presumably will expend a relatively large amount of effort to correct high-confidence errors during the feedback trial, thus causing these errors to have a high probability of being corrected. High-confidence correct responses will be maintained by the feedback trial and thus have a high probability of being remembered. Of course, maintaining responses is easier than correcting errors. Therefore, the perseveration rate of high-confidence correct responses will be higher than the error correction rate for high-confidence incorrect responses, causing the likelihood function to be asymmetrical. Because low-confidence responses receive very little maintenance or error correction effort, the likelihood of correct R3 responses following an initial low-confidence response will be about the same for initially correct responses and initially incorrect responses.

4.2 Empirical Findings

Very few studies have examined the effect of response confidence on the effectiveness of feedback, but those studies have tended to support the servocontrol model of Kulhavy and Stock (Butterfield & Mangels, 2003; Butterfield & Metcalfe, 2001; Kulhavy & Stock, 1989; Kulhavy, Stock, Thornton, & Winston, 1990; Stock, Kulhavy, Pridemore, & Krug, 1992). Webb, et al. (1994) further examined the influence of response confidence on the effect of variations in the timing of feedback in two experiments that simultaneously tested the interference perseveration hypothesis and the servocontrol model of response certitude. Because these experiments demonstrate the strengths and weaknesses of both theories, they are discussed in detail below.

In Experiment 1 of Webb, et al. (1994), re-presentation timing (EOT-I vs. EOT-D, 24-hour delay) was manipulated between-subjects and re-presentation type (feedback vs. no feedback) was manipulated within-subjects. Participants started by taking an initial multiple-choice test over general knowledge questions. For each question, they were shown the question and asked to judge how confident they were that they could answer it correctly. They were then shown the 4 choices and asked to choose the correct response and to make another confidence judgment. After completing the test, participants saw the same questions again at the designated delay. In the no feedback condition, the question was presented again without any of the response alternatives; in the feedback condition, the question and response alternatives were shown with the correct response marked by an asterisk. Seven days after the initial presentation, participants took the same test using the same procedures as had been used on the initial test. Experiment 2 was similar to Experiment 1 except that feedback type was manipulated between-subjects, participants only made confidence judgments on half the items, and the confidence judgments were only made after answering the test questions.

Consistent with the interference-perseveration hypothesis, Webb, et al. (1994, Experiment 1) found that delayed feedback (M $\approx$ .28) lead to better performance on the final test than did immediate feedback (M $\approx$ .24); the size of this effect was comparable with that seen in previous studies that have found evidence of the delay-retention effect. However, in Experiment 2, they did not find any significant differences between the immediate feedback group and the delayed feedback group. The fact that they failed to replicate the main results of Experiment 1 despite the similar procedures indicates, at the very least, that the delay-retention effect is highly sensitive to relatively minor experimental manipulations.

Webb et al.'s (1994) results for the effects of response confidence on study time are similarly inconsistent. Experiment 1 showed the general pattern for study time predicted by the servocontrol theory of response certitude, with high-confidence errors having the longest average study time and high-confidence correct responses having the shortest average study time, regardless of the timing of the feedback. Experiment 2, on the other hand did not show the predicted pattern of study times.

Finally, Webb et al.'s (1994) results for the effects of response confidence on final test performance are also problematic. The shapes of the conditional probability curves for initially-correct items were consistent with the theoretical predictions in both experiments, with the probability of correct response on the final test being a nearly linear function of the confidence in the initial response. However, the shapes of the conditional probability curves for error correction were not consistent with the theoretical predictions in either experiment. Except in the delayed feedback condition of Experiment 2, medium-confidence errors were the most likely to be corrected, and high-confidence errors were no more likely to be corrected than were low-confidence errors.

4.3 Summary

Common sense says that learners' confidence in their previous responses plays a role in the processing of feedback. This supposition is backed up by empirical evidence from a relatively small number of studies. The only theoretical explanation of the impact of confidence on feedback processing that has been offered to date, the servocontrol model of Kulhavy and Stock (1989), is complex and ill-specified and can only explain part of the existent data. Clearly, additional theoretical development is needed.

CHAPTER 5

EXPERIMENTS

We have seen that the effects of feedback timing on semantic learning are poorly understood. There is a large amount of variation in the empirical literature, with some studies finding that immediate feedback leads to better long-term retention, some studies finding that delayed feedback leads to better long-term retention, and other studies failing to find any effect of varying the timing of feedback. Additionally, the current theoretical explanations for the effect of feedback timing cannot fully describe the pattern of findings in the literature. Discrepancy-reduction theories such as operant conditioning theory and connectionist models predict that immediate feedback should always lead to better long-term retention, while the interference perseveration hypothesis—the leading currently accepted theoretical account—predicts the opposite.

In response, I have offered the spacing hypothesis, based on the new theory of disuse (Bjork & Bjork, 1992), as an alternative explanation for the findings in the literature. The spacing hypothesis appears to be able to explain the complete pattern of empirical results seen in the literature using the same theoretical structure that has already been applied to a wide variety of other learning phenomena. Thus, the spacing hypothesis is potentially a more parsimonious and more powerful theoretical account of feedback timing effects than any of the theories that have been advanced to date.

Despite the compelling theoretical arguments for the spacing hypothesis, the empirical evidence is still somewhat weak because the designs used in previous studies of feedback timing effects do not allow for a direct test of the spacing hypothesis. Many of the previous studies have focused on direct applications to classroom learning (see Table

2.1) and, in doing so, have given up the tight control over study time and restudy opportunities that is needed to test the spacing hypothesis. Other studies have controlled these aspects but failed to control other critical variables such as prior knowledge of the to-be-learned material. Additionally, almost all of the feedback timing studies have used multiple-choice tests, introducing a large number of possible confounds and limiting the interpretability of their findings.

This chapter reports the results of a set of three experiments that were designed to provide a direct test of the spacing hypothesis and other theoretical explanations of feedback timing effects on retention of meaningful information. The design of the experiments and the choice of stimuli reflect my desire to test these theories in a well-controlled laboratory setting while also maximizing ecological validity. To this end, these new experiments tested manipulations of spacing, lag, and type of re-presentation (study, test, feedback) for new learning of semantic information. The information to be learned in each of the experiments consisted of trivia facts that participants were likely to find interesting but were unlikely to know prior to beginning the experiment. Also, to avoid the confounds and complications associated with multiple-choice tests, these new experiments used constructed response (i.e., cued recall) tests. Experiment 1 was designed to test the claim underlying the spacing hypothesis that feedback timing effects are empirically equivalent to spacing effects. Experiments 2 and 3 were designed to see whether manipulations based on predictions from the new theory of disuse could produce the complete pattern of results from the feedback timing literature.

5.1 Experiment 1

Experiment 1 was designed to test whether the effect of varying the timing of feedback is empirically equivalent to the effect of varying the timing of pure study and pure test trials. The design directly tests the central claim underlying the spacing hypothesis and also tests the predictions of the new theory of disuse regarding the relative effectiveness of study and test trials. The central questions Experiment 1 was designed to answer are:

- Is providing feedback more effective than providing a restudy or retest opportunity in promoting long-term retention?
- Is the effect of delaying feedback different from spacing effects for restudy and retest trials?

To answer these questions, Experiment 1 directly compared spacing effects (massed vs. spaced study), test-spacing effects (massed vs. spaced tests), and feedback timing effects (immediate feedback vs. delayed feedback) on a cued recall test following a one week retention interval. As detailed in Chapter 3, the new theory of disuse predicts that manipulations that decrease the retrieval strength for a learning event increase the potency of that learning event (Bjork & Bjork, 2006). Based on this idea, I would expect the following pattern of results for Experiment 1.

First, there should be an overall main effect of re-presentation timing—the spaced\delayed conditions should lead to better final test performance on average than the massed\immediate conditions—because the retrieval strength on delayed trials will be lower on average than retrieval strength on massed trials, regardless of the type of re-presentation. This also implies that a spacing effect should be evident for each type of re-presentation. That is, a basic spacing effect should be observed for the restudied items; a test-spacing effect should be observed for the retested items; and a feedback spacing effect should be observed in the feedback conditions. For items that were answered correctly on the initial test, a feedback trial can be considered a restudy trial; it is these items in particular that should show evidence of the feedback spacing effect.

In addition to the spacing effects, the new theory of disuse also predicts that there should be effects of testing (Bjork & Bjork, 1992) and error correction. Because test trials are more effortful than study trials, each of the test conditions should result in better retention than their corresponding study conditions. Similarly, because the feedback conditions incorporate a test trial, final test performance should be higher for the feedback conditions than for the pure study conditions. Due to the different impacts that feedback trials and test trials have on correctly-answered items compared to incorrectly-answered items, final test performance for the feedback conditions should be about the same or somewhat greater than performance for the retest conditions, depending on the efficacy of error correction. Items that are answered correctly on the initial test should benefit more from an additional test trial than from a feedback trial that acts as a restudy trial. However, items that are answered incorrectly on the initial test receive no such benefit in the retest condition but can benefit from the error correction function in the feedback conditions. Thus, the benefit for additional testing in the retest condition can be canceled out or even surpassed by the benefit of error correction in the feedback condition. As described in Chapter 3, the new theory of disuse also predicts that in new learning—such as in this experiment—errors are relatively easy to correct but some errors will not be corrected because of interference effects. These interference effects should be reduced in the delayed feedback condition, thus increasing the probability of correcting errors in the delayed feedback condition relative to the immediate feedback condition.

Unlike the new theory of disuse, the interference perseveration hypothesis does not provide a firm basis for making predictions regarding spacing effects for pure study or retest events; the only predictions for this experiment that the interference perseveration hypothesis can reliably make are for the feedback conditions. As detailed in Chapter 2, the interference perseveration hypothesis predicts that delayed feedback should lead to improved performance on a retention test compared to immediate feedback because the delay reduces interference and increases the probability of error correction (Kulhavy & Anderson, 1972). These basic predictions are identical to the predictions made by the new theory of disuse. However, unlike the new theory of disuse, the interference perseveration hypothesis states that the only effect of feedback on items that were correctly answered is to maintain these items in memory (Kulhavy & Anderson). Thus, the interference perseveration hypothesis predicts that there should not be a feedback spacing effect.

5.1.1 Method

5.1.1.1 Participants

Participants were 88 undergraduate students enrolled in psychology courses at the University of Texas at Arlington who participated for partial course credit. Data from one participant were not properly stored due to an equipment malfunction. Data from participants who failed to complete both experimental sessions ($n = 5$) and who did not follow instructions during the study session ($n = 2$) were excluded from further analysis, leaving usable data from 80 participants to be analyzed.

5.1.1.2 Design

Experiment 1 used a 2 × 3 within-subjects design in which type of re-presentation (restudy, retest, or feedback) was crossed with spacing of representation (immediate/massed or delayed/spaced). Each experimental condition consisted of three trials—an initial study trial followed by 2 study trial, 2 test trials, or a test with a study/feedback trial. Using S to represent a study trial, T to represent a test trial, and a dash to represent a delay of approximately 9.5 min, the six resulting conditions were as follows: immediate restudy (S-SS), immediate retest (S-TT), immediate feedback (S-TS), delayed restudy (S-S-S), delayed retest (S-T-T), and delayed feedback (S-T-S). Immediate feedback was operationally defined as IBI-I feedback, and delayed feedback was operationally defined as IBI-D feedback.

5.1.1.3 Materials

A total of 64 trivia facts were selected from a database of trivia facts that have been normed for base knowledge rate and interestingness in a separate study that was conducted at the University of Texas at Arlington. The trivia facts in the database were drawn from a variety of general knowledge domains, including history, science, sports, famous people, and popular culture. Each trivia fact is in the form of a question and a short answer of one or two words or numbers. For this experiment, facts were selected that were rated as highly interesting by the respondents in the norming study but for which none of the respondents knew the correct answer beforehand. For facts with answers that are proper names of people, the first name was displayed in parenthesis, and the participants were only required to learn the last name. An example of one such fact is: *Who was the only president with a Ph.D.?* The answer, of course, is *(Woodrow) Wilson.*

Four facts were used as buffer items to control for primacy effects, and the other facts were randomly assigned to the experimental conditions such that there were 10 facts in each condition. The assignment of facts to conditions was randomly determined anew for each participant, and the presentation order of the facts was block randomized for each participant.

5.1.1.4 Procedure

The experiment was conducted in a series of two sessions, a 60 min study session and a 30 min test session one week later at the same time of day. Informed consent was obtained orally before the study session began and again before the test session began.

At the beginning of the study session, each participant was seated in front of an eMac computer which was used to present the trivia questions and answers and to record the participant's responses. Participants were given oral instructions by the experimenter while being shown sample displays corresponding to each task on the computer. After the instructions, participants were left alone to complete the experiment proper.

During the study session, each of the trivia facts was presented once in each of three phases. In the initial phase each of the trivia facts was presented in a 12s study trial. In each study trial, the question appeared near the top of the screen in a black 24 point Monaco font, and the answer appeared immediately below the question in the same font but in blue letters. A text box in which the participant was to type his/her response appeared below the space for the correct answer. To ensure that the study and test tasks were as equivalent as possible, participants were instructed to type the correct answer into the text box during study trials as well as during test trials. After 12s had elapsed, the computer presented a blank screen for 500ms before presenting the next trivia fact.

After all the trivia facts were studied once, participants were given a 60s distracter task (a reaction time game) before beginning the second phase. During the second phase, each of the questions in the immediate re-presentation conditions were presented twice, and each of questions in the delayed re-presentation conditions were presented once. For the immediate restudy condition, each fact was presented in 2 back-to-back study trials; for the immediate retest condition, each question was presented in 2 back-to-back test trials; and for the immediate feedback condition, each question was presented in a test trial followed immediately by a study/feedback trial. The test trials were identical to the study trials, with the obvious exception that the correct answer was not displayed.

Before the third phase, participants played the reaction time game again for 155s, and then all the questions in the delayed re-presentation conditions were presented again, with the final trial being a study trial (in the delayed restudy and delayed feedback conditions) or a test trial (in the delayed retest condition). The presentation order in all three phases was block randomized to ensure that different presentation orders were used in each of the three study phases. The average lag time between the initial trial and the second trial for each fact was approximately 9.5 min for all conditions, and the average lag time between the second trial and the third trial was also approximately 9.5 minutes for the delayed re-presentation conditions.

In the test session, participants were given a cued recall test over all 64 questions they had studied the previous week. With the exception of the 4 buffer questions, the presentation order for the test questions was randomly determined anew for each participant. During the test, each of the trivia questions was presented using the same procedure that was used for the test trials in the study session. After completing the test, each participant was debriefed, thanked for their participation, and released.

5.1.2 Results

5.1.2.1 Scoring

Participants' responses on each trial were scored using both strict and liberal scoring criteria. For strict scoring, participants' responses to each question had to lexically match the originally displayed correct answer in order to be counted as a correct response. There were two exceptions to this rule: Correct numerical answers were accepted whether the participant responded using words or digits, and spaces in compound words such as *starfish* were ignored. Items for which a participant did not type any response, for which they only typed one letter, or for which they clearly indicated they did not know the answer—such as by typing "I don't remember"—were counted as blank responses. All other responses longer than one letter were counted as incorrect responses. Thus, analyses conducted using strict scoring counted misspellings, typographic errors, and variations in tense or number as incorrect responses.

For liberal scoring, responses that semantically matched the originally displayed correct answer were counted as correct. Variations in spelling, part of speech, tense, or number were counted as correct as long they were consistent with the meaning of the correct answer. Proper names spelled phonetically were also counted as correct responses, as were responses that were incomplete but could clearly be completed only one way—such as when the last letter or two was missing. Scoring of blank responses used the same rules as were used for strict scoring. All responses not scored as correct or blank were counted as incorrect.

All planned analyses were conducted using both sets of scoring criteria, and the basic pattern of results was similar for both sets. Because the liberal scoring criteria more accurately record semantic learning, only the results using liberal scoring are reported in this paper. Alpha was set to .05 for all inferential analyses.

5.1.2.2 Final Test Performance

A 2 × 3 repeated measures ANOVA revealed significant main effects of re-presentation timing, F(1,79) = 4.15, MSE = 0.031, p = .05, and re-presentation type F(2,158) = 4.82, MSE = 0.031, $p < .01$, on final test performance. There was no significant interaction between timing and type of re-presentation, F(2,158) = 1.51, MSE = 0.015, $p >$ .10.

As can be seen in Figure 5.1, delayed re-presentation improved retention overall (M = .72, SE = .01) relative to immediate representation (M = .69, SE = .02). This main effect was due primarily to the study and test conditions, as revealed by planned simple effect comparisons. Delaying restudy (M = .70, SE = .03) increased retention relative to immediate restudy (M = .67, SE = .03), although this effect did not quite reach the traditional level of statistical significance, t(79) = 1.87, p = .07. Delaying retest (M = .71, SE = .02) significantly increased retention compared to immediate retesting (M = .66, SE = .02), t(79) = 2.18, p = .03. However, there was no significant effect of delaying feedback, t(79) = 0.29, $p >$ .10.

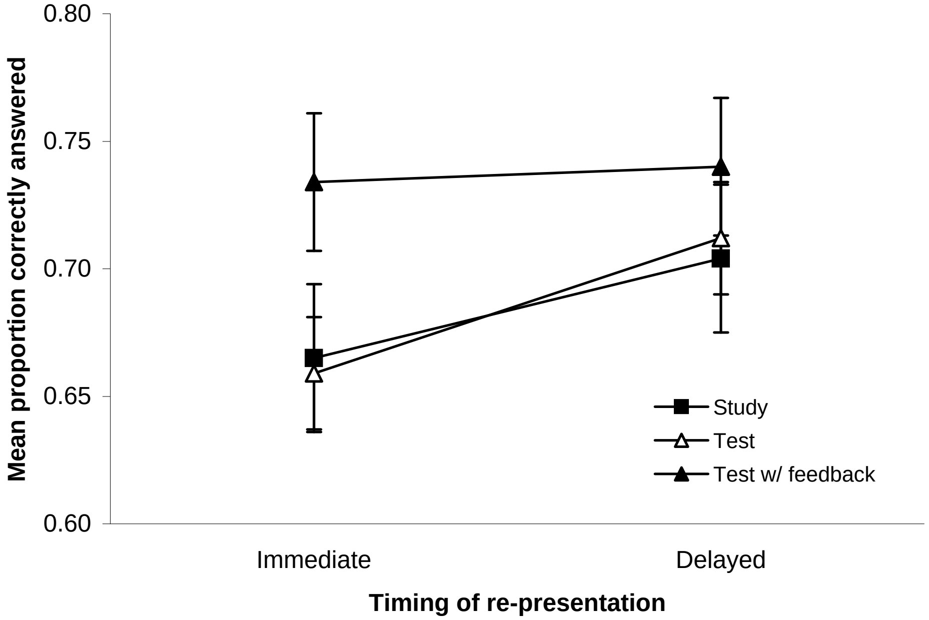

Figure 5.1. Mean correct recall on the final retention test in Experiment 1. Error bars represent standard errors.

Planned comparisons for the simple effects of re-presentation type revealed that there were no significant differences on final test performance between the restudy and retest conditions for immediate re-presentation, $t(79)$ = -.22, p > .10, or for delayed representation, $t(79) = 0.31$, p > .10. However, the feedback condition did exhibit better final test performance than the restudy condition both for immediate re-presentation, $t(79) = 3.54$, p < .01, and for delayed re-presentation, $t(79) = 2.22$, p = .03. Comparisons of the feedback condition to the test condition showed that the feedback condition also exhibited significantly better final test performance than the test condition for immediate re-presentation, $t(79) = 2.69$, p < .01, but the difference for delayed re-presentation did not approach statistical significance, $t(79) = 1.20$, p > .10.

5.1.2.3 Conditional Analysis

As described above, each response on the initial test and the final retention test was scored as either a correct response (R), an incorrect response (W), or a blank (B). Using these three categories, I constructed conditional probability trees for each of the test and feedback conditions in order to test the detailed predictions regarding the differential effects of delay, testing, and feedback on items correctly answered on the initial test compared to items that were not correctly answered on the initial test.

Examining the conditional probability trees for the test conditions (Figure 5.2), we can see that the response probabilities are nearly identical on the first test; this is to be expected because there are no differences in the experimental conditions until after the first test. There is also very little difference in the conditional probabilities on the final test for initially incorrect and blank items. As predicted by the new theory of disuse, the effect of delay for a retest is confined to the items that were initially answered correctly. Here we can see evidence for the test-spacing effect, with the probability of remembering an initially correct response being substantially higher in the delayed retest condition (M = .84, SE = .01) than in the immediate retest condition (M = .78, SE = .02).

Turning to the conditional probability trees for the feedback conditions (Figure 5.3), we see the expected similarity in response patterns on the first test. However, we do not see the predicted feedback spacing effect for the initially correct responses; the conditional probability of remembering an initially correct response is not substantially higher in the delayed feedback condition (M = .88, SE = .01) than in the immediate feedback condition (M = .86, SE = .01). Nor do we see any difference in the probability of correcting initially incorrect responses. We see only small differences in the probability of correcting initial blank responses. In fact, comparing the probability trees for the feedback conditions (Figure 5.3) to those for the test conditions (Figure 5.2), we see that

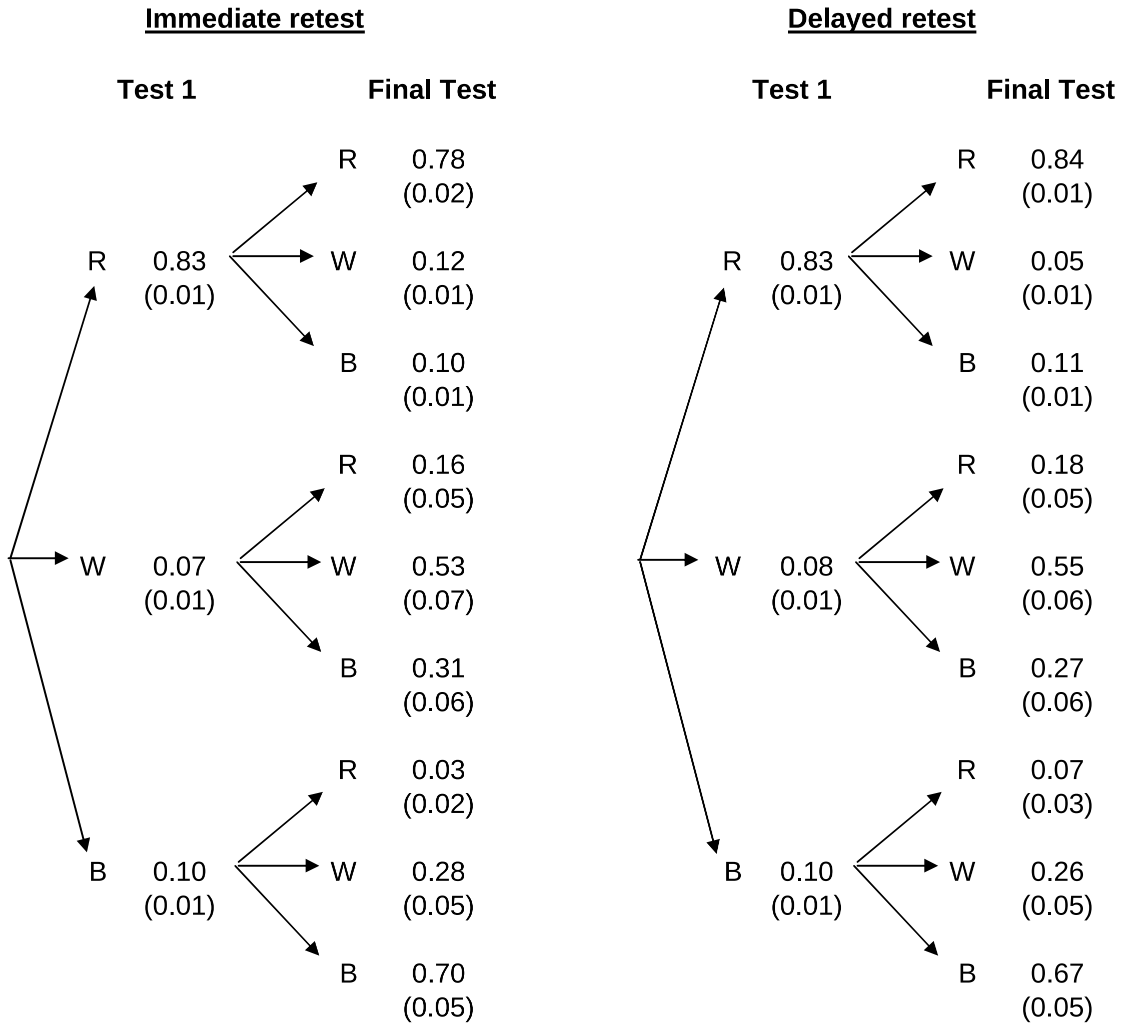

Figure 5.2. Conditional probability trees for the retest conditions in Experiment 1. R = right (i.e., correct response). W = wrong (i.e., incorrect response). B = blank (i.e. no response provided). Numbers in parentheses are standard errors calculated across all available observations.

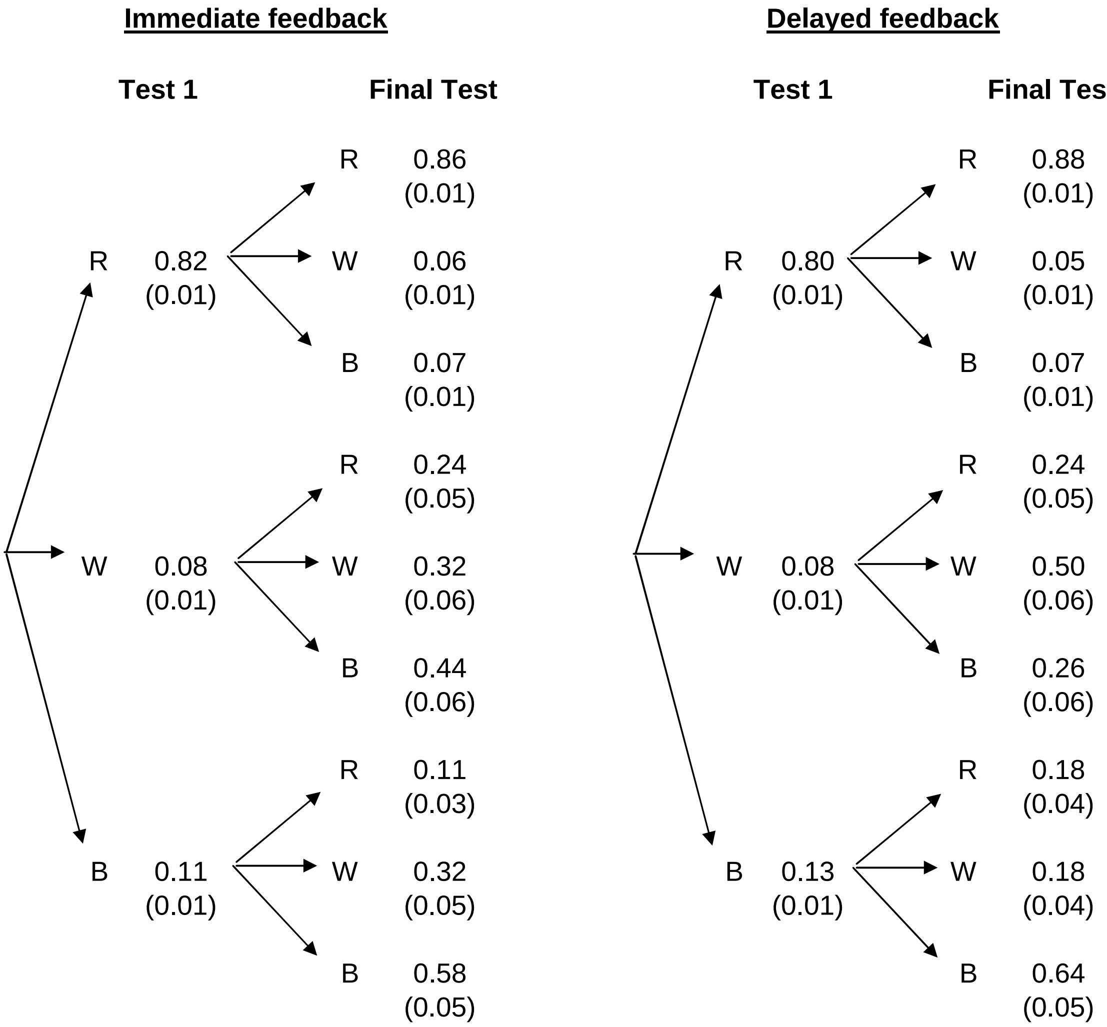

Figure 5.3. Conditional probability trees for the feedback conditions in Experiment 1. R = right (i.e., correct response). W = wrong (i.e., incorrect response). B = blank (i.e. no response provided). Numbers in parentheses are standard errors calculated across all available observations.

the error correction rate following feedback was not much higher than the error correction rate when no feedback was provided. However, feedback substantially improved retention of initially correct responses more than an additional test, especially for the immediate re-presentation conditions.

5.1.3 Discussion

The results of Experiment 1 are generally consistent with the predictions of the new theory of disuse. A main effect of spacing was observed as predicted, as were the predicted simple spacing effects for restudying and retesting. Also consistent with the new theory of disuse, the test-spacing effect was revealed to be due to an increased probability of remembering initially correct responses following a delayed retest trial.

The failure to observe the predicted testing effect may be due to the relatively short initial study-test interval. A 9.5 min study-test interval was chosen in order to allow a moderate amount of forgetting to occur between the initial study trial and the first test trial. According to the new theory of disuse, the testing effect is obtained when retrieval is more difficult on a test trial than on a corresponding study trial. However, the conditional analysis showed that on average the correct responses were remembered 83% of the time on the first test. Given this high level of performance on the first test, it appears that the average retrievability of items on the first test trial in the test condition was not much different from the perfect retrievability of items in the study condition. If this was the case, then the lack of a testing effect is not surprising and is consistent with the new theory of disuse. However, the lack of differences in overall final test retention and in error correction probabilities between the immediate feedback condition and the delayed feedback condition is somewhat surprising, especially when one considers that both the new theory of disuse and the interference perseveration hypothesis predict these effects.

5.2 Experiment 2

Experiment 2 was designed to test in more detail the theoretical predictions of the new theory of disuse and the interference perseveration hypothesis regarding feedback timing effects. In particular, Experiment 2 used experimental manipulations suggested by predictions from the spacing hypothesis to attempt to replicate the complete pattern of feedback timing results from the literature within a single experiment. Experiment 2 was also designed to investigate the effect of re-processing opportunities after receiving feedback and to examine differential effects of feedback on correct vs. incorrect responses. Additionally, confidence ratings were collected in order to test whether confidence in initial test responses has any impact on the effect of feedback in the various conditions.

In Experiment 2, all eight experimental conditions included an initial study trial, a test trial, and a feedback trial, with feedback provided immediately after the test trial in half of the conditions and after a delay of about 10 minutes in the other half of the conditions. This manipulation of feedback timing and the manipulations of event sequences described below were chosen in order to be able to replicate the basic designs of previous feedback timing studies. Some of these previous studies have limited participants' exposure to the stimuli to the basic study-test-feedback cycle; others have allowed participants restudy opportunities after the feedback was provided; and a few have even retested the material at a short retention interval following the presentation of feedback. Therefore, Experiment 2 included a pair of conditions which consisted of the standard study-test-feedback trials with feedback presented either immediately or at a delay (the no re-presentation conditions), a pair of conditions that included a restudy trial a short time after the feedback trial (the restudy conditions), and a similar pair of conditions that included a retest trial shortly after the feedback trial (the retest conditions). In each of these six conditions, the initial study-test lag was about the same as was used in Experiment 1. In order to test the effect of variations in the initial study-test lag on the effect of feedback, an additional pair of no re-presentation conditions with twice the initial study-test lag (the long S-T lag conditions) was included.

Based on the spacing hypothesis and the new theory of disuse, I expected to find a main effect of feedback timing, with the delayed feedback conditions leading to better retention on average than the immediate feedback conditions (i.e., a feedback-spacing effect). I also expected that there would be simple effects of feedback timing, but I anticipated that these effects might be modulated by the type of re-presentation. Because restudy and retest events increase storage and retrieval strengths, they should dilute the feedback-spacing effect. Therefore, I expected the simple effect of feedback timing to be smaller for these conditions than for the no re-presentation condition. The additional increments in storage and retrieval strength should also increase the overall retention in the restudy and retest conditions relative to the no re-presentation condition.

One of the potential criticisms of the new theory of disuse is that at first glance it does not seem to comport with the finding that immediate feedback can actually lead to better retention than delayed feedback in some circumstances. However, as described in Chapter 3, many of the studies that have found an advantage for immediate feedback have allowed participants restudy opportunities in the immediate feedback conditions that were not allowed in the delayed feedback condition. The comparison of the delayed no-representation condition to the immediate restudy condition in Experiment 2 is conceptually similar to these studies. In this comparison, spacing of restudy opportunities has been equated, and the only difference is whether or not a feedback trial immediately follows the test trial. Because the spacing is now equivalent, the new theory of disuse would no longer predict an effect of delaying feedback. Immediate feedback following a correct retrieval on the first test can be considered a massed study trial; therefore, this should have a relatively small impact on retention of correct responses. The immediate feedback can also correct initial errors, especially errors of omission, leading to the possibility of an improvement in learning and retention. The probability of this improvement is a direct function of the number of errors on the initial test that could be corrected. If few errors are made on the first test, then there should be very little, if any, difference in performance between the two conditions. However, if a sufficient number of errors are made on the first test, then immediate feedback followed by a spaced restudy should lead to better performance on the final retention test than delayed feedback with no restudy.

Unlike the new theory of disuse, the interference perseveration hypothesis is not general enough to distinguish between the various conditions in Experiment 2. Specifically, it does not provide a solid basis for making predictions concerning potential interactions between feedback timing and re-presentation of information following the feedback trial. The only predictions that can be made based on the interference perseveration hypothesis are the same as those made for the feedback conditions in Experiment 1. Delayed feedback should lead to an improvement in final retention test performance relative to immediate feedback due to an increased probability of error correction, but there should be no differences in the probability of remembering an initially correct response between these conditions.

Confidence judgments were collected on the test trials in Experiment 2 in order to test the predictions of the servocontrol model of response certitude (Kulhavy & Stock, 1989) and to attempt to replicate the finding by Butterfield & Metcalfe (2001) that high confidence errors are hypercorrected relative to low confidence errors. As described in Chapter 4, the servocontrol model stipulates that feedback is driven by a discrepancy reduction mechanism, with the efficacy of a feedback trial being driven by how confident the participant was when they responded to the preceding test trial and by the degree of match between the participant's test response and the provided feedback. The level of discrepancy can be calculated by multiplying the confidence judgment by +1 for incorrect responses or by -1 for correct responses. It is this discrepancy that is thought to drive participants' motivation, and thus their behavior in response to the feedback (Kulhavy & Stock). Kulhavy and Stock found a strong log-linear relationship between the calculated discrepancy and the amount of time participants spent studying the feedback. They also found evidence of the V-shaped conditional probability curve predicted by the servocontrol model. In their study, the probability of initially correct answers perseverating to the final test was inversely related to discrepancy, and the probability of correcting an initial error was directly related to discrepancy. According to the servo-control model, this same pattern should be observed in Experiment 2.

5.2.1 Method

5.2.1.1 Participants

Participants were 121 undergraduate students enrolled in psychology courses at the University of Texas at Arlington who participated for partial course credit. Data from 7 subjects were not properly stored due to equipment malfunctions and experimenter error. Data from participants who failed to complete both experimental sessions ($n = 9$) or who did not follow instructions during the study session ($n = 2$) were excluded from further analysis, leaving usable data from 103 participants to be analyzed.

5.2.1.2 Design

Experiment 2 used a quasi-factorial within-subjects design with eight conditions. The design consisted of a 2 × 3 factorial manipulation of feedback timing (immediate vs. delayed) and type of re-presentation after the feedback trial (none, restudy, or retest) along with an overlapping 2 × 2 factorial manipulation of feedback timing (immediate vs. delayed) and initial study-test interval (short vs. long). Using S to represent a study trial, T to represent a test trial, and a dash to represent a delay of 8-10 min, the eight resulting conditions were as follows: immediate feedback with no re-presentation (S-TS), immediate feedback with restudy (S-TS-S), immediate feedback with retest (S-TS-T), delayed feedback with no re-presentation (S-T-S), delayed feedback with restudy (S-T-S-S), delayed feedback with retest (S-T-S-T), immediate feedback with long study-test lag (S–TS), and delayed feedback with long study-test lag (S–T-S). The first six conditions comprised the 2 × 3 portion of the design, while the first two and last two conditions comprised the 2 × 2 portion of the design.

5.2.1.3 Materials

A total of 76 trivia facts were used. These trivia facts included the 64 facts from Experiment 1 plus 12 additional facts selected from the same database using the same criteria that were used in Experiment 1. Twelve facts were used as buffer items to control for primacy effects and to help maintain spacing of the trials. The other facts were randomly assigned to the experimental conditions such that there were 8 facts in each condition. The assignment of facts to conditions was randomly determined anew for each participant, and the presentation order of the facts was randomized for each participant.

5.2.1.4 Procedure

The basic procedure for Experiment 2 was the same as the procedure that was used for Experiment 1, with three notable exceptions. First, in order to allow for a total of eight within-subject conditions rather than six, the time for each trial was reduced from 12s to 10s[1]. The shorter initial study time was also intended to reduce the strength

[1] Although 10s does not seem like much time to respond to these stimuli, a preliminary examination of response times from Experiment 1 indicated that participants needed less than 6s on average to answer questions on the test trials.

of initial encoding and allow for more forgetting between the initial study trial and the first test trial. Second, because of the complexity of the design, a scheduling algorithm similar to the one used by Pashler, Zarow, and Triplett (2003) was used to schedule the trials for each trivia fact during the study session (see Appendix A for the scheduling algorithm), and the reaction-time game distracter task was not used. Third, on each test trial in the study session, participants were asked to judge how confident they were that their answer was correct using a 7-point Likert scale, with 1 representing "Not at all confident" and 7 representing "Very confident".

As in Experiment 1, participants returned to the lab one week after the study session for a final test session. The procedures for the test session were similar to those used in Experiment 1. During the test, each of the 76 trivia questions was presented using the same procedure that was used for the test trials in the study session, including asking participants to provide a confidence rating. After completing the test, each participant was debriefed, thanked for their participation, and released.

5.2.2 Results

Because of the quasi-factorial nature of the experimental design, a set of analyses in which the 6 short lag conditions were compared to each other was conducted, and a parallel set of analyses in which the short lag no re-presentation conditions were compared to the long lag conditions was conducted. For each of these factorial sets, analyses of initial test performance and of final test performance were conducted using repeated measures ANOVAs with planned follow-up comparisons. Conditional analyses were also conducted in order to examine the probabilities of correct answer perseveration, error perseveration, and error correction. Alpha was set to .05 for all analyses. Participants' responses were scored using the same liberal scoring criteria as was used in Experiment 1.

5.2.2.1 Comparison of Short Lag Conditions to Long Lag Conditions

5.2.2.1.1 Performance on the First Test

A 2 × 2 repeated measured ANOVA revealed a significant main effect of initial study-test lag on the proportion of correct responses on the first test, $F(1{,}102) = 8.38$, $MSE = 0.029$, $p < .01$, with participants remembering less in the long lag conditions ($M = .68$, $SE = .02$) than the short lag conditions ($M = .72$, $SE = .01$). There were no significant differences in performance due to feedback timing, nor was there a significant interaction between study-test lag and feedback timing, both Fs < 1. Thus, the study-test lag manipulation was successful in inducing forgetting between the initial study trial and the first test trial, although the amount of forgetting was quite a bit less than I was hoping to see.

5.2.2.1.2 Performance on the Final Retention Test

Difference scores were used to measure retention because of the expected baseline differences in performance on the first test. Difference scores were calculated for each item by subtracting the initial test score from the final test score. Thus, items that were answered correctly on both tests or were not answered correctly on both tests each received a score of 0; items that were not answered correctly on the final test despite having been answered correctly on the first test each received a score of -1; and items that were not answered correctly at first but then were answered correctly on the final test each received a score of +1. These item-level difference scores were then used to perform the planned analyses of main effects, simple effects, and interactions.

A 2 × 2 repeated measures ANOVA revealed significant main effects of study-test lag, $F(1{,}102) = 18.53$, $MSE = 0.018$, $p < .01$, and feedback delay, $F(1{,}102) = 20.90$, $MSE = 0.026$, $p < .01$, on retention. These main effects were qualified by a significant interaction, $F(1{,}102) = 11.01$, $MSE = 0.023$, $p < .01$ (see Figure 5.4). Providing imme-

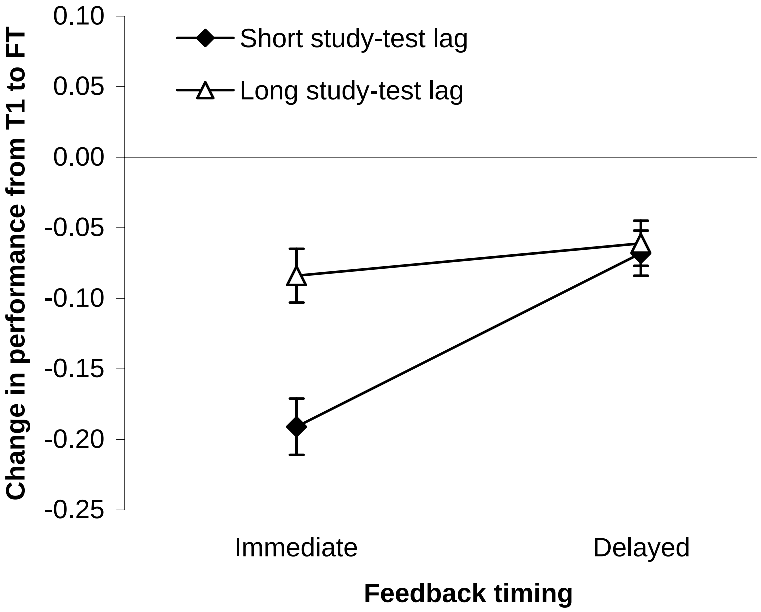

Figure 5.4. Mean change in performance between the initial test (T1) and the final retention test (FT) for the no re-presentation conditions in Experiment 2. Error bars represent standard errors.

diate feedback was more effective after a long study-test lag than after a short study-test lag, $t(102) = 5.02$, $p < .01$, but there was no difference in the efficacy of delayed feedback for these conditions, $t(102) = 0.39$, $p > .10$.

5.2.2.1.3 Conditional Analysis

Conditional probability trees for the no re-presentation conditions are shown in Figure 5.5. An examination of the response probabilities on the first test shows that participants' lower rate of correct responding (R) in the long lag conditions than in the short lag conditions was accompanied by a higher rate of null responses (B), $t(102) = 3.51$, $p < .01$, but not by a higher rate of incorrect responses (W), $t(102) = 0.17$, $p > .10$. Because the participant's had not been exposed to these stimuli prior to the experiment, any incorrect responses on the first test can be assumed to be guesses. Therefore, the longer lag made the correct responses less available on the first test but did not change the guessing rate.

Turning to the final test, we can see evidence of the spacing effect predicted by the new theory of disuse for items that were correctly answered on the initial test. As was previously discussed, a feedback trial for these items can be considered as a restudy opportunity. After a short study-test lag, delaying feedback by 8 min (i.e. spaced study) dramatically increased the probability of remembering the correct answer a week later compared to providing immediate feedback ($\Delta M = .15$). A less dramatic increase is also seen when comparing the long lag immediate feedback condition to the long lag delayed feedback condition ($\Delta M = .04$). Interestingly, delaying feedback seems to have increased error correction probabilities for the short lag conditions and decreased error correction probabilities in the long lag conditions, although these apparent effects are small and not statistically reliable.

5.2.2.2 Comparison of No Re-presentation, Restudy, and Retest Conditions

5.2.2.2.1 Performance on the First Test

Because the 6 short lag conditions used identical procedures on the trials up to and including the first test, there was no theoretical reason to expect any differences in performance on this test. Nevertheless, I examined initial test performance using a 2 × 3 repeated measures ANOVA in order to check for any possible baseline variations. The ANOVA revealed a significant difference in the baseline proportion of correct responses on the initial test between the immediate and delayed feedback conditions, $F(1,102) = 4.29$, $MSE = 0.019$, $p = .04$. These small baseline differences between the conditions in initial test performance are likely the result of random non-experimental variations. As can be seen in Table 5.1, the means for the delayed restudy and delayed retest condition are slightly lower than the means for the other short lag conditions. Notably, the reduction in

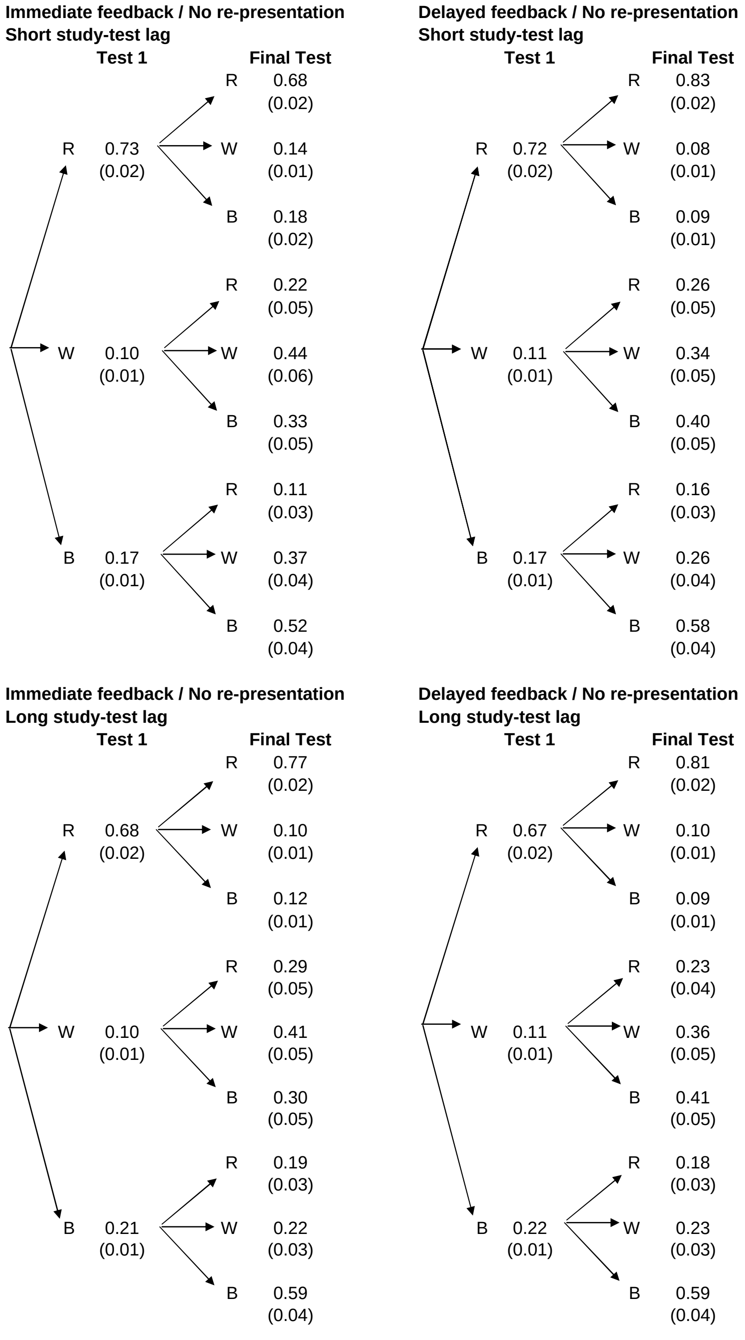

Figure 5.5. Conditional probability trees for the no re-presentation conditions in Experiment 2. R = right (i.e., correct response). W = wrong (i.e., incorrect response). B = blank (i.e. no response provided). Numbers in parentheses are standard errors calculated across all available observations.

Table 5.1. Mean Proportion of Answers Correctly Recalled on the Initial Test in Experiment 2

Study-test lag	Type of re-presentation	Feedback timing	
		Immediate	Delayed
Long	None	.68	.67
Short	None	.73	.72
Short	Restudy	.71	.68
Short	Retest	.71	.68

presentation time from Experiment 1 was successful in reducing the overall performance on the initial test by about 10 points as compared to the corresponding conditions in Experiment 1.

5.2.2.2.2 Performance on the Final Retention Test

Because of the differences in baselines, performance on the final test was measured in the form of difference scores—as was done in the comparison of short lag to long lag conditions—rather than as a simple proportion on items answered correctly. This also has the advantage of making the two sets of analyses comparable.

Figure 5.6 shows the mean difference scores for the 3 short study-test lag conditions in Experiment 2. A 2 × 3 repeated measures ANOVA revealed significant main effects of feedback timing, $F(1{,}102) = 38.24$, $MSE = 0.019$, $p < .01$, and type of re-presentation following feedback, $F(2{,}204) = 25.99$, $MSE = 0.028$, $p < .01$, on the change in performance from the first test to the final retention test, as well as a significant interaction of feedback timing and re-presentation type, $F(2{,}204) = 4.73$, $MSE = 0.025$, $p < .01$. As can be seen in Figure 5.6, delaying feedback improved retention overall compared to providing immediate feedback. This effect was statistically significant when there was no re-presentation of the material, $t(102) = 5.27$, $p < .01$, and when material was retested,

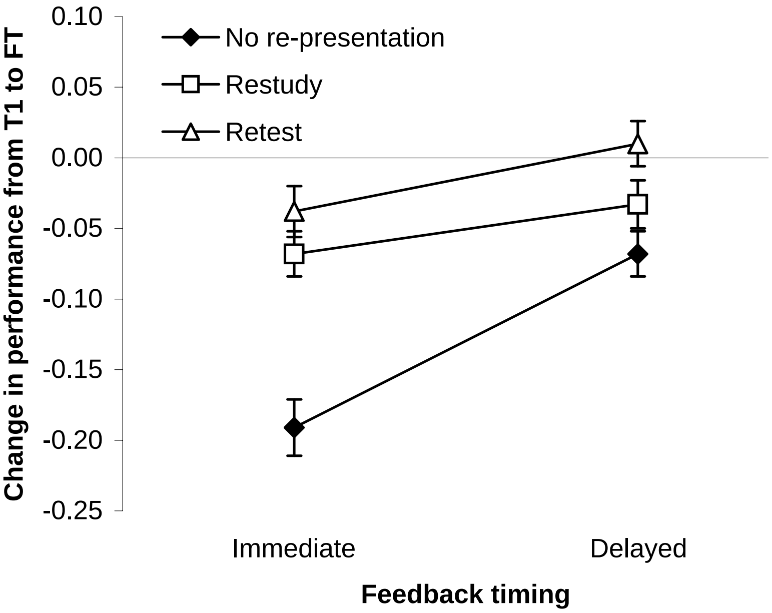

Figure 5.6. Mean change in performance between the initial test (T1) and the final retention test (FT) for the no re-presentation, restudy, and retest conditions in Experiment 2. Error bars represent standard errors.

$t(102) = 2.31$, $p = .02$, and it approached significance when material was restudied, $t(102) = 1.88$, $p = .06$. The interaction of feedback timing and re-presentation type was as predicted by the new theory of disuse, with the effect of delaying feedback being moderated by restudy and retest trials. Also as predicted by the new theory of disuse, restudying following feedback improved retention compared to no re-presentation, $t(102) = 4.85$, $p < .01$, and retesting improved retention compared to restudying, $t(102) = 2.22$, $p = .03$. Finally, there was no difference between the immediate feedback with restudy condition and the delayed feedback with no re-presentation condition (both Ms = -.07)—again, as predicted by the new theory of disuse.

5.2.2.2.3 Conditional Analysis

Conditional probability trees for the restudy and retest conditions are shown in Figure 5.7. First, we can see that the response patterns on the first test are quite similar to the pattern for the short lag no re-presentation condition (compare Figure 5.7 to the top of Figure 5.5). Second, we can see that compared to the no re-presentation condition, restudying or retesting an item increased the probability of remembering a correct response following immediate feedback (ΔM = .12 and .14, respectively). We can also see evidence for a much smaller spacing effect for the restudy (ΔM = .03) and retest conditions (ΔM = .05) than was seen in the no re-presentation condition (ΔM = .15). This smaller spacing effect appears to be due mostly to the increased probability of remembering correct responses when items are restudied or retested after immediate feedback, as can be seen by comparing $P(R_2|R_1)$ for immediate feedback in the no-representation condition (M = .68) to the same probability in the restudy and retest conditions (Ms = .80 and .82, respectively) .

A quick examination of the error correction probabilities, $P(R|W)$ and $P(R|B)$, shows that restudying and retesting items after the presentation of feedback increases the effectiveness of the feedback in correcting errors relative to not seeing the items again. Compared to the no re-presentation condition, restudying increased the probability of correcting both previous incorrect responses and previous null responses (ΔM = .06 and .09, respectively). Retesting an item increased the probabilities of error correction even further. Compared to restudying, retesting increased the probability of correction by .07 for a previous incorrect response and for a previous null response. Providing immediate feedback followed by a retest after a short delay was particularly effective in correcting errors, $P(R|W) = .44$.

5.2.2.3 Confidence Judgments and Error Perseveration

Because the timing of all trials was controlled in Experiment 2, the relationship between feedback study time and response confidence that is predicted by the servocontrol model of Kulhavy and Stock (1989) could not be tested in Experiment 2. The fact

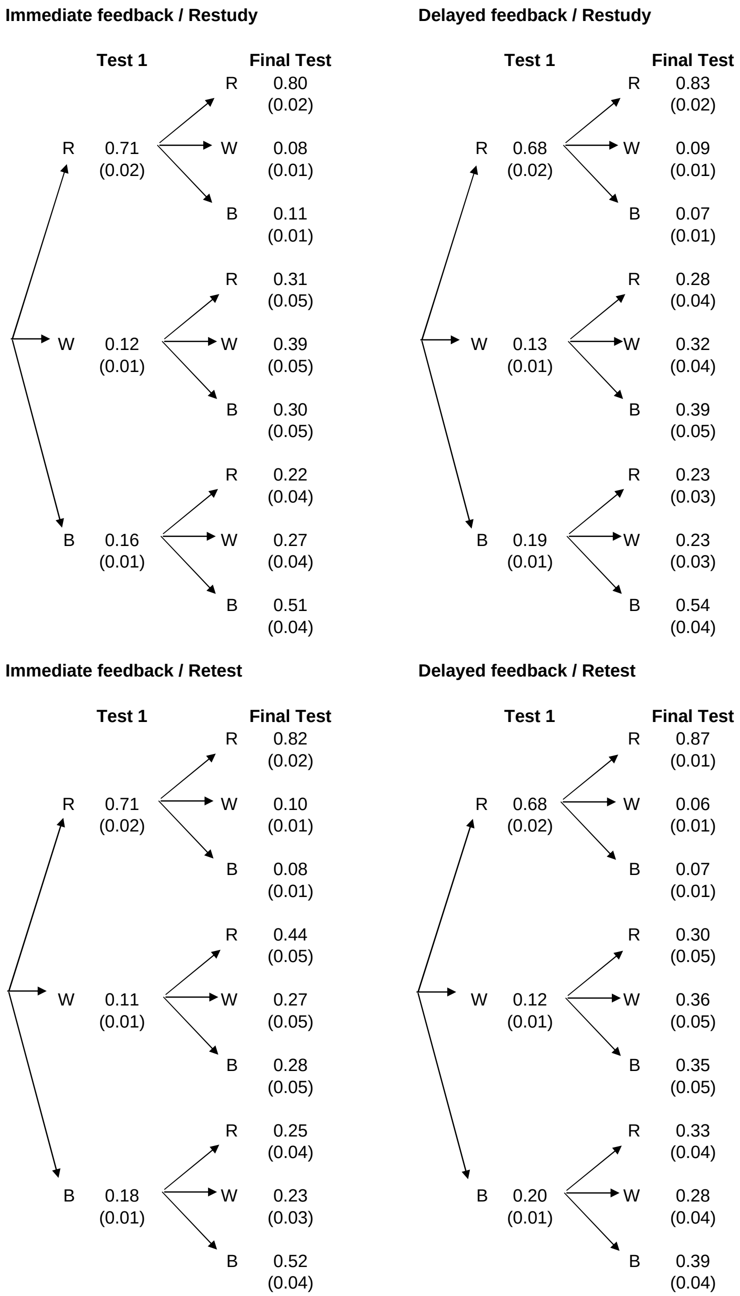

Figure 5.7. Conditional probability trees for the restudy and retest conditions in Experiment 2. R = right (i.e., correct response). W = wrong (i.e., incorrect response). B = blank (i.e. no response provided). Numbers in parentheses are standard errors calculated across all available observations.

that the test trials were timed also impaired the collection of confidence ratings because participants could not be forced to provide a confidence rating before proceeding to the next trial. On average, participants gave confidence judgments on 77.4% of test trials. These issues will be dealt with in Experiment 3.

A discrepancy index was calculated for each item that was given a confidence judgment on the initial test by multiplying the confidence judgment by +1 if the initial test response was incorrect or by -1 if the initial test response was correct. Figure 5.8 shows the proportion of items correctly answered on the final retention test as a function of this discrepancy index for the immediate feedback conditions and for the delayed feedback conditions in Experiment 2. Only half of the V-shaped pattern predicted by the servocontrol model can be seen. Consistent with the model, there was a direct relationship between initial confidence and correct answer perseveration. However, this finding is consistent with the general idea that confidence ratings can be considered as an index of the ease of retrieval and thus, does not distinguish the servocontrol model from other theoretical perspectives. As can be seen from the right side of Figure 5.8, there was no evidence for the unique prediction of the servocontrol model that high confidence errors are more likely to be corrected. Figure 5.9 shows additional evidence of this: there was no substantive relationship between confidence ratings and error perseveration rates. Figure 5.9 also shows that, contrary to the predictions of the servocontrol model and the interference perseveration hypothesis, delaying feedback did not reduce the probability of an initial error being repeated on a later retention test.

5.2.3 Discussion

The results of Experiment 2 are consistent with the predictions of the new theory of disuse but are not consistent with predictions from the interference perseveration hypothesis. As predicted by both theories, retention over a 7-day interval was better

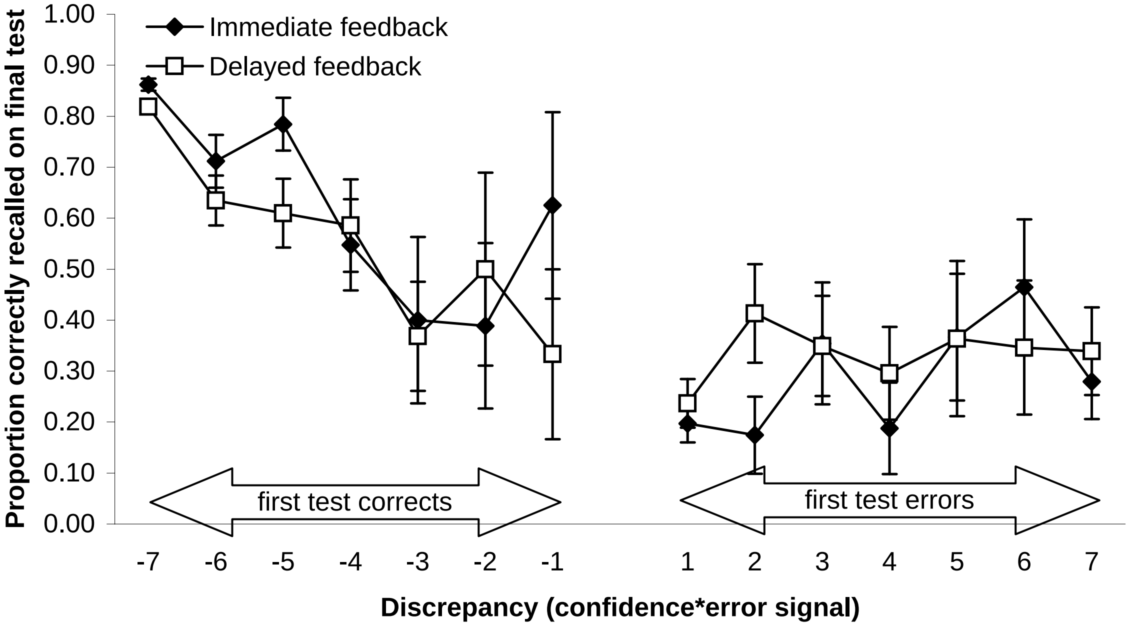

Figure 5.8. Probability of correct responding on the final retention test as a function of feedback discrepancy in Experiment 2. Error bars represent standard errors.

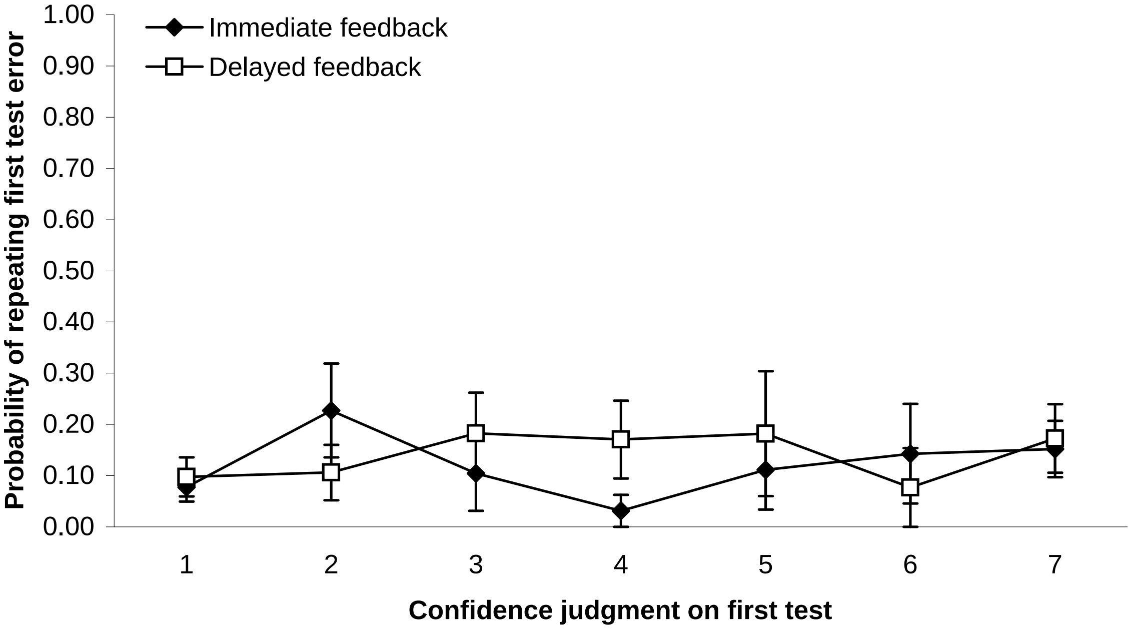

Figure 5.9. Probability of error perseveration from the initial test to the final retention test as as a function of feedback discrepancy in Experiment 2. Error bars represent standard errors.

when delayed feedback was provided than when immediate feedback was provided. There was no evidence for a difference in error correction probabilities as was predicted by the interference perseveration hypothesis. Instead, the difference in retention was caused primarily by the increased probability of remembering an initially correct response; this is clear evidence of the feedback spacing effect predicted by the new theory of disuse and a clear violation of the predictions of the interference perseveration hypothesis.

The new theory of disuse predicted that the feedback spacing effect would be moderated by restudying and by retesting after the feedback trial; this predicted pattern of overall retention was obtained, with retest > restudy > no-representation. Varying the initial study-test lag also had a dramatic impact on the relative effects of feedback timing. This finding was not predicted but is consistent with the new theory of disuse as described in Chapter 3.

Experiment 2 was successful in replicating the finding that allowing participants to restudy or retest themselves after an immediate feedback trial but not after a delayed feedback trial can lead to a null effect or to an observed benefit for immediate feedback. This result was predicted by the new theory of disuse and provides a ready explanation for many of the studies that have found that immediate feedback leads to better retention than does delayed feedback.

The analysis of confidence judgments provided conflicting evidence regarding the servocontrol model. There was no evidence for the hypercorrection of high confidence initial errors, but there was evidence of confidence effects on initially correct responses. These findings must be interpreted with caution however, as participants only gave confidence ratings on 77% of test trials.

5.3 Experiment 3

I elected to control the timing at every stage in Experiments 1 and 2 in order to rule out potential confounds between study time and the experimental manipulations. Nevertheless, Experiment 2 might be criticized on the basis that participants' did not have sufficient time to answer the test questions or to process the feedback. Also, because the timing is controlled at every stage, I was not able to examine potentially interesting relationships between study time and participants' responses nor was I be able to collect confidence judgments for every item. Experiment 3 alleviates these concerns by making all trials after the initial study trial self-paced rather than timed and by requiring confidence judgments on every test trial. In all other respects, Experiment 3 was identical to Experiment 2.

5.3.1 Method

5.3.1.1 Participants

Participants were 126 undergraduate students enrolled in psychology courses at the University of Texas at Arlington who participated for partial course credit. Data from 1 subject were not properly stored due to experimenter error. Data from participants who failed to complete both experimental sessions ($n = 12$) and who did not follow instructions during the study session ($n = 6$) were excluded from further analysis, leaving usable data from 107 participants to be analyzed.

5.3.1.2 Design, Materials, and Procedure

Experiment 3 used the same design, materials, and procedure as were used in Experiment 2 with the following three exceptions. First, confidence ratings were required rather than just requested on all test trials. Second, all trials after the initial 10s study trial were self-paced rather than timed. After typing a response (and making a confidence judgment on the test trials), participants clicked a “Done” button to complete the trial and move on to the next item. To ensure that participants did not spend too much time on any single item, an audio warning to hurry was played if the participant took longer than 30s to complete a trial; this warning was then repeated every 5s until the participant completed the trial. Third, the number of stimuli per condition was reduced from 8 to 7 in order to ensure that all participants would finish in the allotted time, even if they took longer than 10s per trial on average. Each of these changes applied to the initial study session and to the final test session one week later.

5.3.2 Results

Participants’ responses were scored using the same liberal scoring criteria that were used in Experiments 1 and 2. As in Experiment 2, final test performance was measured using difference scores. Each of the planned analyses that were conducted for Experiment 2 were repeated for Experiment 3. Because the time for each trial was under participant control rather than experimenter control, response times were examined as well. Alpha was set to .05 for all analyses.

5.3.2.1 Comparison of Short Lag Conditions to Long Lag Conditions

5.3.2.1.1 Response Times

The mean time participants took to respond to a test item was 7.9s. A 2 × 2 repeated measures ANOVA revealed no significant differences in test response time between any experimental conditions, all Fs $<$ 2 and all ps $>$ .10. The mean time participants spent studying feedback was 4.8s per item. A 2 × 2 repeated measures ANOVA revealed a significant effect of feedback timing on study time, $F(1{,}107) = 20.72$, $MSE = 999{,}529$, $p < .01$, with participants spending an average of 4.5s studying each item on immediate feedback trials and 5.0s studying each item on delayed feedback trials. The effect of initial study-test lag on study time approached statistical significance, $F(1{,}107) = 3.64$, $MSE = 1{,}061{,}106$, $p = .06$. Participants studied each feedback message an average of 4.9s in the short lag conditions and 4.7s in the long lag conditions. There was no significant interaction between feedback timing and study-test lag, $F(1{,}107) = 2.61$, $MSE = 746{,}861$, $p > .10$. On average, participants saw each question in the no re-presentation conditions for a total of 23s during the study phase of the experiment.

5.3.2.1.2 Performance on the First Test

A 2 × 2 repeated measures ANOVA revealed a significant main effect of initial study-test lag on the proportion of correct responses on the first test, $F(1{,}107) = 3.96$, $MSE = 0.027$, $p = .05$, with participants remembering less in the long lag conditions ($M = .67$, $SE = .01$) than the short lag conditions ($M = .71$, $SE = .01$). There were no significant differences in performance due to feedback timing, $F < 1$, nor was there a significant interaction between study-test lag and feedback timing, $F = 1.19$, $MSE = .023$, $p > .10$. Thus, the study-test lag manipulation was successful in inducing forgetting between the initial study trial and the first test trial, and this portion of the results from Experiment 2 was replicated.

5.3.2.1.3 Performance on the Final Retention Test

As in Experiment 2, difference scores were used to measure retention because of the expected baseline differences in performance on the first test. A 2 × 2 repeated measures ANOVA revealed a significant main effect of study-test lag, $F(1{,}107) = 15.95$, $MSE = 0.025$, $p < .01$, with the long lag conditions leading to better retention than the short lag conditions (see Figure 5.10), and a significant main effect of feedback timing, $F(1{,}107) = 5.94$, $MSE = 0.026$, $p = .02$, with delayed feedback leading to better retention than

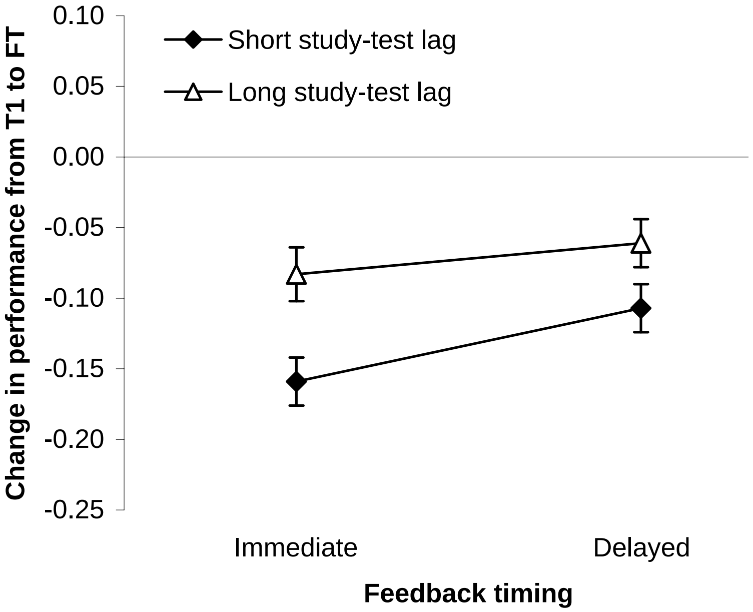

Figure 5.10. Mean change in performance between the initial test (T1) and the final retention test (FT) for the no re-presentation conditions in Experiment 3. Error bars represent standard errors.

immediate feedback. Unlike Experiment 2, there was no significant interaction between study-test lag and feedback timing, $F < 1$. Comparing Figure 5.10 to Figure 5.4, we can see that performance in the long lag conditions were almost identical in the two experiments; the only differences were in the short lag conditions. It appears that giving participants control over the timing of trials increased the effectiveness of immediate feedback and decreased the effectiveness of delayed feedback in these conditions.

5.3.2.1.4 Conditional Analysis

Conditional probability trees for the no re-presentation conditions are shown in Figure 5.11. An examination of the response probabilities on the first test shows that, just as in Experiment 2, participants' lower rate of correct responding (R) in the long lag conditions than in the short lag conditions was accompanied by a higher rate of null responses (B), $t(107) = 2.59$, $p = .01$, but not by a higher rate of incorrect responses (W),

t(107) = -0.45, $p > .10$. As in Experiment 2, the longer lag made the correct responses less available on the first test but did not change the guessing rate. The spacing effect for items that were correctly answered on the initial test that was observed in Experiment 2 was also observed in Experiment 3. Compared to providing immediate feedback, delaying feedback increased the probability of remembering an initially correct response in both the short lag conditions ($\Delta M = .08$) and the long lag conditions ($\Delta M = .04$). The only portion of the conditional analysis from Experiment 2 that was not replicated in Experiment 3 was the pattern of error correction probabilities. When participants were allowed to control the timing of the test and feedback trials, there were no apparent differences between the effects of immediate and delayed feedback on error correction or error perseveration.

5.3.2.2 Comparison of No Re-presentation, Restudy, and Retest Conditions

5.3.2.2.1 Response Times

The mean time participants took to respond to a test item was 7.9s. A 2 × 3 repeated measures ANOVA revealed no significant differences in test response time between any experimental conditions, all Fs < 1. The mean time participants spent studying feedback was 4.9s per item. A 2 × 3 repeated measures ANOVA revealed a significant effect of feedback timing on study time, F(1,107) = 18.30, MSE = 1670144, $p < .01$, with participants spending an average of 4.6s studying each immediate feedback message and 5.1s studying each delayed feedback message. There was no significant effect of re-presentation on study time, $F < 1$, and no significant interaction, , $F < 1$. In the restudy conditions, participants restudied each item for 4.4s on average, and there was no significant difference between the immediate feedback condition and the delayed feedback condition, t(107) = -1.41, $p > .10$. In the retest conditions, participants took an average of 6.6s to respond to a retested item, with participants spending significantly less time

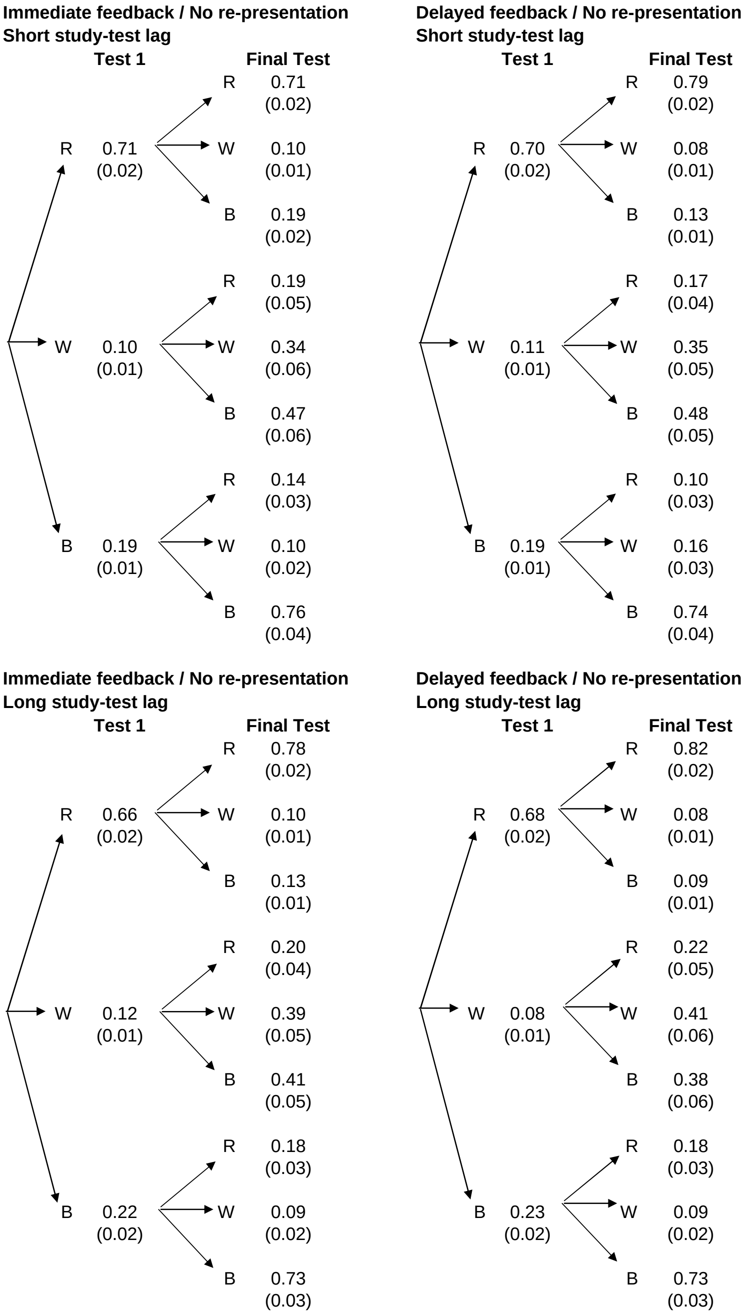

Figure 5.11. Conditional probability trees for the no re-presentation conditions in Experiment 3. R = right (i.e., correct response). W = wrong (i.e., incorrect response). B = blank (i.e. no response provided). Numbers in parentheses are standard errors calculated across all available observations.

to respond to an item tested after immediate feedback (M = 6.4s) than after delayed feedback (M = 6.8s), t(107) = -2.88, p < .01. During the study phase of the experiment, participants saw each question for an average of 23s in the no re-presentation conditions, 27s in the restudy conditions, and 30s in the retest condition. Importantly, when participants were allowed to control the timing, they studied each item for substantially less time on average than the 10s per trial that was used in Experiment 2.

5.3.2.2.2 Performance on the First Test

A 2 × 3 repeated measures ANOVA was used to check for any possible baseline variations in performance on the first test. As would be expected, there were no significant differences due to feedback timing, $F < 1$, type of re-presentation, $F < 1$, nor was there a signficant interaction between timing and presentation type, F(2,214) = 1.60, MSE = 0.029, p > .10. The mean proportion of correct responses on the first test (M = .71) was the same as was observed in Experiment 2.

5.3.2.2.3 Performance on the Final Retention Test

Even though there were no baseline differences, final test retention was measured as the difference between performance on the first test and performance on the final test. This measurement was used to enable comparisons to the previous analyses performed for Experiments 2 and 3.

Figure 5.12 shows the mean difference scores for the short study-test lag conditions in Experiment 3. A 2 × 3 repeated measures ANOVA revealed a significant main effect of re-presentation following feedback, F(2,214) = 27.98, MSE = 0.026, p < .01, on the change in performance from the first test to the final retention test. The main effect of feedback timing, F(1,107) = 2.90, MSE = 0.033, p = .09, and the interaction of feedback timing and re-presentation type, F(2,214) = 2.90, MSE = 0.023, p = .06,

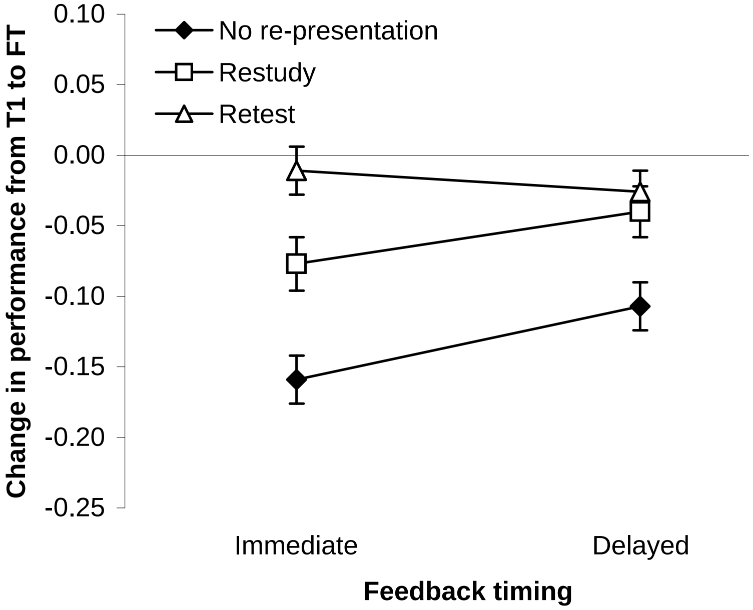

Figure 5.12. Mean change in performance between the initial test (T1) and the final retention test (FT) for the no re-presentation, restudy, and retest conditions in Experiment 3. Error bars represent standard errors.

approached but did not reach the level of statistical significance. As can be seen in Figure 5.12, delaying feedback significantly improved retention compared to providing immediate feedback for the no re-presentation conditions, $t(107) = 2.14$, $p = .03$, and marginally improved retention in the restudy conditions, $t(107) = 1.74$, $p = .08$. Feedback timing did not have a significantly effect on retention in the retest conditions, $t(107) =$ -0.75, $p > .10$.

5.3.2.2.4 Conditional Analysis

The conditional analysis of the no re-presentation, restudy, and retest conditions shows further replication of the pattern of findings from Experiment 2. First, the patterns of first test responses in each of the short lag conditions in Experiment 3 are almost identical to each other (Figure 5.13 and top of Figure 5.11) and to the corresponding conditions from Experiment 2 (Figure 5.7 and top of Figure 5.5). This was expected because there were no procedural differences between any of these conditions until after the first test. Second, the probability of remembering a correct response following immediate feedback was higher when the item was restudied ($M = .80$) or retested ($M = .85$) than when the item was not re-presented ($M = .71$), just as in Experiment 2. Third, there was a large feedback spacing effect for initially correct responses in the no re-presentation conditions ($\Delta M = .08$) and a much smaller feedback spacing effect for the restudy conditions ($\Delta M = .03$), also as was observed in Experiment 2. Fourth, the effects of restudy and retest trials on error correction probabilities were similar to those seen in Experiment 2. Compared to the no re-presentation condition, restudying increased the probability of correcting both previous incorrect responses and previous null responses (ΔMs $= .12$), as did retesting (ΔMs $= .12$ and $.15$, respectively).

There were only two discrepancies between the conditional analyses of Experiments 2 and 3. Unlike in Experiment 2, there was no evidence for a feedback spacing effect in the retest conditions in Experiment 3. Also, Experiment 3 did not replicate the finding from Experiment 2 that the the most effective way to increase the efficacy of error correction is to retest the items a short time after providing immediate feedback.

5.3.2.3 Confidence Judgments and Error Perseveration

The basic findings from Experiment 2 regarding the effects of confidence on feedback were also replicated in Experiment 3. First, as shown in Figure 5.14, high confidence correct responses were more likely to be remembered a week later than were low confidence correct responses and there were no substantive differences in the patterns for immediate and delayed feedback[2]. We can also see that there is no substantive evidence for the predicted hypercorrection of high confidence errors.

[2] Extreme caution should be taken in examining individual data points on this graph because not every cell mean is based on data from the entire set of participants. For example, although it appears

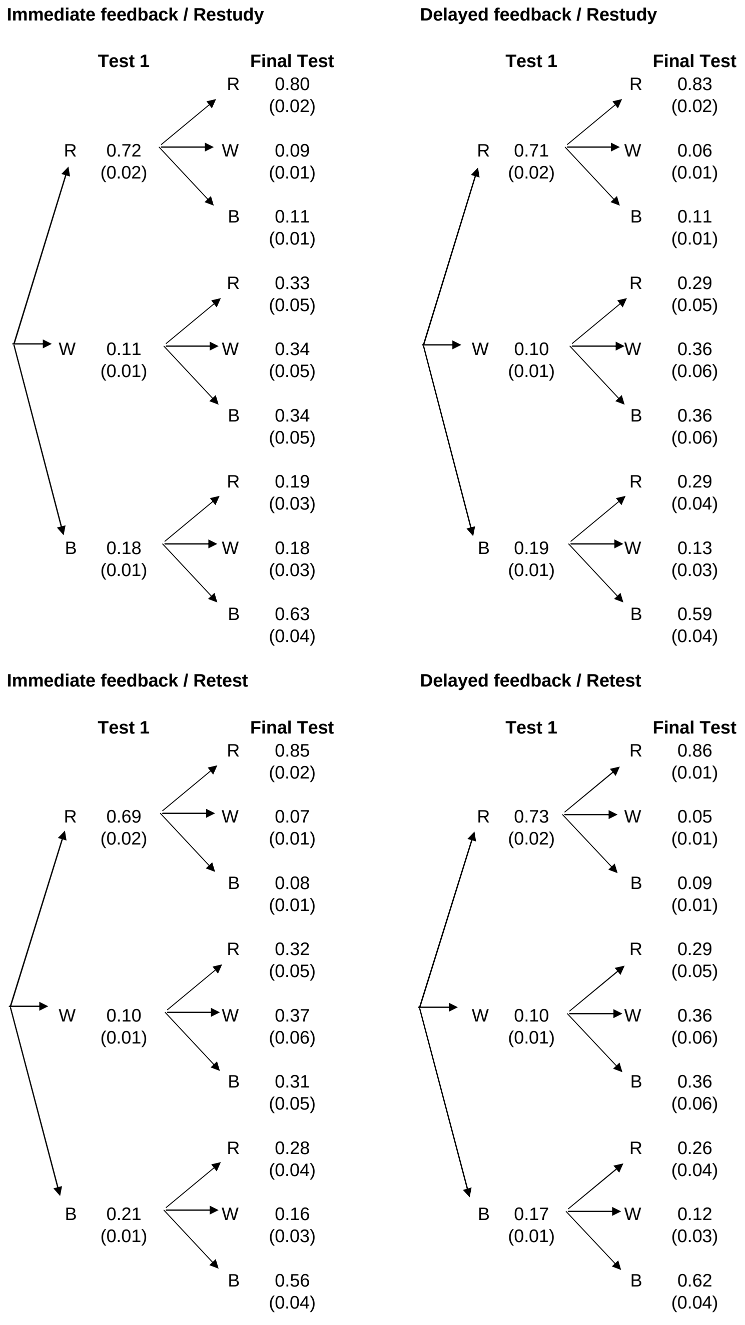

Figure 5.13. Conditional probability trees for the restudy and retest conditions in Experiment 3. R = right (i.e., correct response). W = wrong (i.e., incorrect response). B = blank (i.e. no response provided). Numbers in parentheses are standard errors calculated across all available observations.

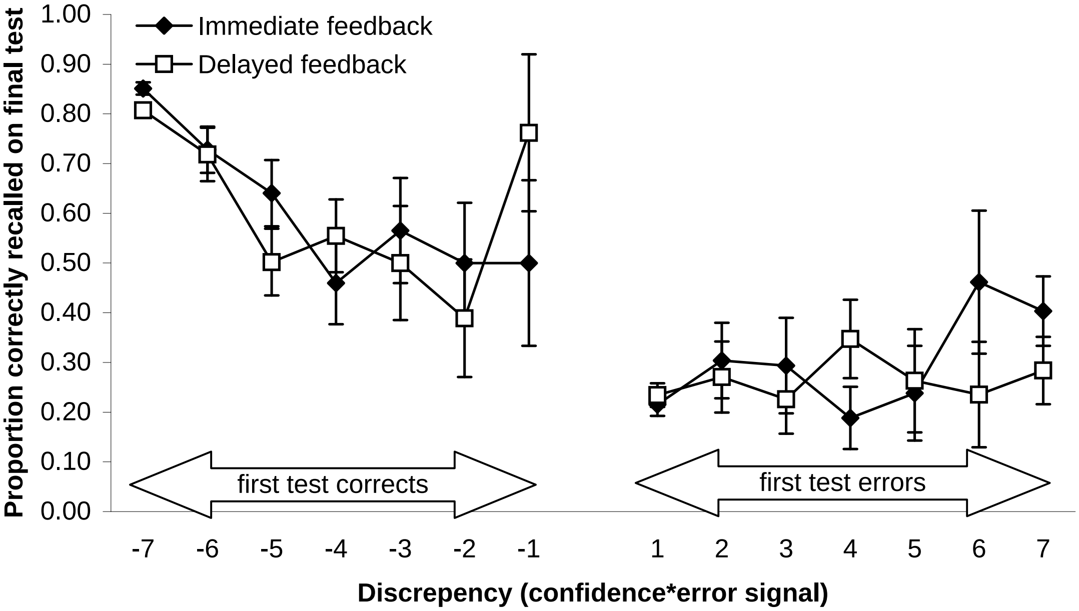

Figure 5.14. Probability of correct responding on the final retention test as a function of feedback discrepancy in Experiment 3. Error bars represent standard errors.

Turning to the effects of confidence on feedback study time, we can see limited evidence for the trend predicted by the servocontrol model (Figure 5.15). Participants spent the least amount of time studying feedback for high confidence correct responses (low discrepancy items), and feedback study time generally increased with the discrepancy value. However, the trend is not as regular as would be expected from the servocontrol model (R^2 = .06), nor is the slope as steep as would be expected. Kulhavy and Stock (1989) reported slopes of .06 and .07 based on a $\log_{10}$ transformation of study times (in seconds); a similar analysis collapsing across all conditions in Experiment 3 only yields a slope of .01.

that delayed feedback was highly effective in reinforcing the lowest confidence correct responses (the bump in the graph for discrepancy = -1), only 10 participants provided data for this cell mean.

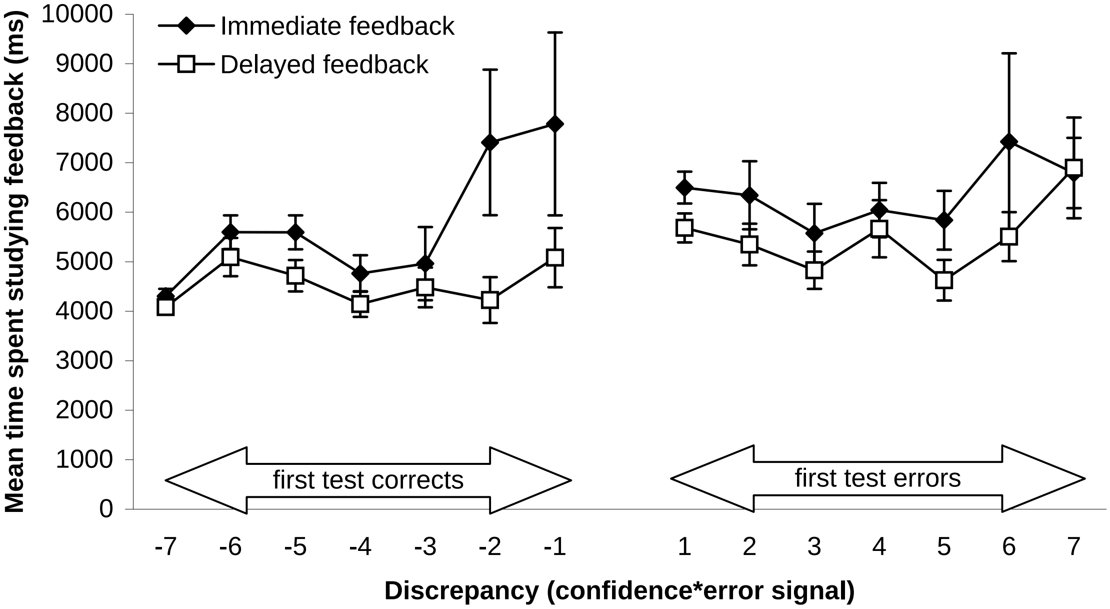

Figure 5.15. Mean feedback study time as a function of feedback discrepancy in Experiment 3. Error bars represent standard errors.

5.3.3 Discussion

Experiment 3 nicely replicated the major findings from Experiment 2. In general, retention over a 7-day interval was was better for delayed feedback than for immediate feedback, whether the amount of time participants spent studying the feedback was controlled by the experimenter (Experiment 2) or the participant (Experiment 3). Conditional analyses of response probabilities revealed that this superiority was caused primarily by a feedback-spacing effect for the initial correct responses, as predicted by the new theory of disuse but contrary to predictions from the interference perseveration hypothesis. Additionally, there was no evidence in either experiment of higher error perseveration rates in the immediate feedback condition than in the delayed feedback condition, a central prediction of the interference perseveration hypothesis.

Experiment 3 provided an even stronger replication than Experiment 2 of the finding that allowing participants to restudy or retest themselves after an immediate feedback trial but not after a delayed feedback trial can result in a reversal of the delay-retention effect. The participants retained more information in the immediate feedback/restudy condition and in immediate feedback/retest condition than in the delayed feedback/no re-presentation condition.

5.4 Summary

Until now, no single theory has been able to explain the the complete pattern of findings in the feedback timing literature. I have proposed that the new theory of disuse (Bjork & Bjork, 1992) might be able to explain this pattern, have described how the theory can be applied to the feedback timing paradigm, and have tested the predictions of the theory in a series of three experiments. The results of these experiments are broadly consistent with the predictions of the new theory of disuse and provide clear evidence of the deficiencies of the major competing theories, the interference perseveration hypothesis (Kulhavy & Anderson, 1972) and the more recent servocontrol model of response certitude in which it has been embedded (Kulhavy & Stock, 1989). In particular, Experiments 2 and 3 show that the new theory of disuse can be extended to the feedback timing paradigm and that each of the three findings in the feedback timing literature—superiority of delayed feedback, superiority of immediate feedback, and null effects—can be produced within a single experiment using theoretically motivated manipulations. In short, the new theory of disuse does appear to be able to explain the effects of feedback timing on semantic learning.

APPENDIX A

SCHEDULING ALGORITHM FOR EXPERIMENTS 2 AND 3

The scheduling algorithm used in Experiments 2 and 3 operates as follows:

1. Set the number of stimuli and number of presentations (either 3 or 4) for each condition.
2. Calculate lags as number of items between presentations based on nominal lag times. The nominal initial study-test lag time was 10.7 min for the short lag conditions and 21.4 min for the long lag conditions. The nominal test-feedback lag for the delayed feedback conditions was 8.0 min, and the nominal feedback-restudy and feedback-retest lag for the re-presentation conditions was 8.0 min.
3. Randomly assign stimuli to conditions. To do this, the algorithm sets up an array of empty serial position slots. It then sequences through the stimuli, assigns each presentation of the stimulus to a serial position slot as described below:
    (a) Assign the initial study trial to the first empty serial position slot.

    (b) Assign the initial test trial to a serial position such that it occurs after all the initial study trials have been completed and in a randomly selected empty serial position slot such that the actual study-test lag is within 20% of the nominal study-test lag of the condition to which the stimulus item has been assigned.

    (c) Assign the feedback trial to a serial position slot:

      i. If the stimulus item is in an immediate feedback condition and the serial position slot immediately following the assigned test trial is empty, then schedule the feedback trial in that serial position slot.

      ii. If the stimulus item is in an immediate feedback condition and the serial position slot immediately following the assigned test trial is not empty, then shift the items occupying the slot and all subsequent slots right one slot and assign the feedback trial to now empty serial position slot immediately after the test trial.

iii. If the stimulus item is in a delayed feedback condition and the serial position slot immediately following the assigned test trial is empty.

iv. If delayed FB, then assign the feedback trial to a randomly selected empty serial position slot such that the actual test-feedback lag is within 20% of the nominal test-feedback lag of the condition to which the stimulus item has been assigned.

(d) If the stimulus item is in a condition that includes a restudy or retest trial, then assign the restudy or retest trial to a randomly selected empty serial position slot such that the actual feedback-restudy or feedback-retest lag is within 20% of the nominal lag of the condition to which the stimulus item has been assigned.

4. Fill the unused serial position slots with filler items.
5. Calculate average lags for each condition and compare to nominal lags. If the average lags are not within 30% of the nominal lags, reassign the stimuli to conditions by repeating step 3.

REFERENCES

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (p. 237-313). San Diego, CA: Academic Press.

Angell, G. W. (1949). The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. *Journal of Educational Research*, 42, 391-394.

Bahrick, H. P., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 344-349.

Beck, F. W., & Lindsey, J. D. (1979). Effects of immediate information feedback and delayed information feedback on delayed retention. *Journal of Educational Research*, 72, 283-284.

Beeson, R. O. (1973). Immediate knowledge of results and test performance. *Journal of Educational Research*, 66, 224-226.

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 2: From learning processes to cognitive processes* (p. 35-67). Hillsdale, NJ: Lawrence Erlbaum.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse. In L.-G. Nilsson & N. Ohta (Eds.), *Memory and society: Psychological perspectives* (p. 109-133). New York: Psychology Press.

Brackbill, Y., Bravos, A., & Starr, R. H. (1962). Delay-improved retention of a difficult task. *Journal of Comparative and Physiological Psychology*, 55, 947-952.

Brackbill, Y., Isaacs, R. B., & Smelkinson, N. (1962). Delay of reinforcement and the retention of unfamiliar, meaningless material. *Psychological Reports*, 11, 553-554.

Brackbill, Y., & Kappy, M. S. (1962). Delay of reinforcement and retention. *Journal of Comparative and Physiological Psychology*, 55, 14-18.

Brosvic, G. M., Epstein, M. L., Cook, M. J., & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom. *Psychological Record*, 55, 401-418.

Brosvic, G. M., Epstein, M. L., Dihoff, R. E., & Cook, M. J. (2006). Acquisition and retention of Esperanto: The case for error correction and immediate feedback. *Psychological Record*, 56, 205-218.

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

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

Buzhardt, J., & Semb, G. B. (2002). Item-by-item versus end-of-test feedback in a computer-based PSI course. *Journal of Behavioral Education*, 11, 89-104.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19, 619-636.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354-380.

Clariana, R. B. (2000). A connectionist model of instructional feedback effects. *Twenty-third Annual Proceedings of Selected Research Papers from the Annual Convention of the Association for Educational Communications and Technology*, 23, 23-26.

Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, 48, 5-21.

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

Dean, R. S., Garabedian, A. A., & Yekovich, F. R. (1983). The effect of modality shifts on proactive interference in long-term memory. *Contemporary Educational Psychology*, 8, 28-45.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1, 309-330.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition* (*2nd ed.*) (p. 317-344). San Diego, CA: Academic Press.

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *Psychological Record*, 53, 533-548.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *Psychological Record*, 54, 207-231.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2005). Adjunctive role for immediate feedback in the acquisition and retention of mathematical fact series by elementary school students classified with mild mental retardation. *Psychological Record*, 55, 39-66.

Ebbinghaus, H. (1885, 1913). *Memory*: *A contribution to experimental psychology* (H. A. Ruger & E. Clara, Trans.). New York: Teachers College, Columbia University.

English, R. A., & Kinzer, J. R. (1966). The effect of immediate and delayed feedback on retention of subject matter. *Psychology in the Schools*, 3, 143-147.

Epstein, M. L., & Brosvic, G. M. (2002a). Immediate feedback assessment technique: Multiple-choice test that 'behaves' like an essay examination. *Psychological Reports*, 90, 226.

Epstein, M. L., & Brosvic, G. M. (2002b). Students prefer the immediate feedback assessment technique. *Psychological Reports*, 90, 1136-1138.

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports*, 88, 889-894.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., et al. (2002). Immediate Feedback Assessment Technique promotes learning and corrects inaccurate first responses. *Psychological Record*, 52, 187-201.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, 62, 369-377.

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

Haynes, C. R. (1974). Delayed feedback and perseveration of interference. *Psychological Reports*, 35, 246.

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

Hull, C. L. (1943). *Principles of behavior*: *An introduction to behavior theory.* Oxford, England: Appleton-Century.

Hull, C. L. (1952). *A behavior system*: *An introduction to behavior theory concerning the individual organism.* New Haven, CT: Yale University Press.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, 30, 138-149.

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

Kincaid, J. P., & Wickens, D. D. (1970). Temporal gradient of release from proactive inhibition. *Journal of Experimental Psychology*, 86, 313-316.

Kulhavy, R. W., & Anderson, R. C. (1972) Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, 63,505-512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, 1, 279-308.

Kulhavy, R. W., Stock, W. A., Thornton, N. E., & Winston, K. S. (1990). Response feedback, certitude and learning from text. *British Journal of Educational Psychology*, 60, 161-170.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, 58, 79-97.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Skykes (Eds.), *Practical aspects of memory* (p. 625-632). London: Academic Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252-271.

Mayer, R. E. (2005). Introduction to multimedia learning. In R. E. Mayer (Ed.), *The cambridge handbook of multimedia learning* (p. 1-16). New York: Cambridge University Press.

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 609-622.

More, A. J. (1969). Delay of feedback and the acquisition and retention of verbal materials in the classroom. *Journal of Educational Psychology*, 60, 339-342.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology (2nd ed.)* (p. 745-783). Mahwah, NJ: Lawrence Erlbaum.

Neill, W. T., & Mathis, K. M. (1998). Transfer-inappropriate processing: Negative priming and related phenomena. In D. L. Medin (Ed.), *The psychology of* *learning and motivation: Advances in research and theory, Vol.* 38. (p. 1-44). San Diego, CA: Academic Press.

Newman, M. I., Williams, R. G., & Hiller, J. H. (1974). Delay of information feedback in an applied setting: Effects on initially learned and unlearned items. *Journal of Experimental Education*, 42, 55-59.

O'Day, E. F., Kulhavy, R. W., Anderson, W., & Malczynski, R. J. (1971). *Programmed instruction: Techniques and trends.* East Norwalk, CT: Appleton-Century-Crofts.

Paige, D. D. (1966). Learning while testing. *Journal of Educational Research*, 59, 276-277.

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

Peeck, J. (1979). Effects of differential feedback on the answering of two types of questions by fifth-and sixth-graders. *British Journal of Educational Psychology*, 49, 87-92.

Peterson, L. R., Hillner, K., & Saltzman, D. (1962). Supplementary report: Time between pairings and short-term retention. *Journal of Experimental Psychology*, 64, 550-551.

Peterson, L. R., Saltzman, D., Hillner, K., & Land, V. (1962). Recency and frequency in paired-associate learning. *Journal of Experimental Psychology*, 63, 396-403.

Petrusic, W. M., & Dillon, R. F. (1972). Proactive interference in short-term recognition and recall memory. *Journal of Experimental Psychology*, 95, 412-418.

Phye, G. D., & Andre, T. (1989). Delayed retention effect: Attention, perseveration, or both? *Contemporary Educational Psychology*, 14, 173-185.

Phye, G. D., & Baller, W. (1970). Verbal retention as a function of the informativeness and delay of informative feedback: A replication. *Journal of Educational Psychology*, 61, 380-381.

Phye, G. D., Gugliemella, J., & Sola, J. (1976). Effects of delayed retention on multiple-choice test performance. *Contemporary Educational Psychology*, 1, 26-36.

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, 1, 19-40.

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology: Interdisciplinary and Applied*, 29, 417-447.

Pressey, S. L. (1963). Teaching machine (and learning theory) crisis. *Journal of Applied Psychology*, 47, 1-6.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 93-134.

Rankin, R. J., & Trepper, T. (1978). Retention and delay of feedback in a computer-assisted instructional task. *Journal of Experimental Education*, 46, 67-70.

Renner, K. E. (1964). Delay of reinforcement: A historical review. *Psychological Bulletin*, 61, 341-361.

Robin, A. L. (1978). The timing of feedback in personalized instruction. *Journal of Personalized Instruction*, 3, 81-88.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210.

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

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

Saltzman, I. J. (1951). Delay of reward and human verbal learning. *Journal of Experimental Psychology*, 41, 437-439.

Sassenrath, J. M., & Yonge, G. D. (1968). Delayed information feedback, feedback cues, retention set, and delayed retention. *Journal of Educational Psychology*, 59, 69-73.

Sassenrath, J. M., & Yonge, G. D. (1969). Effects of delayed information feedback and feedback cues in learning on delayed retention. *Journal of Educational Psychology*, 60, 174-177.

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

Skinner, B. F. (1938). *The behavior of organisms*: *An experimental analysis*. Oxford, England: Appleton-Century.

Skinner, B. F. (1957). *Verbal behavior*. East Norwalk, CT: Appleton-Century-Crofts.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, 30, 641-656.

Stern, D. P. (2004, December). *The round earth and Christopher Columbus.* (Retrieved June 29, 2006 from http://www.phy6.org/stargaze/Scolumb.htm).

Stock, W. A., Kulhavy, R. W., Pridemore, D. R., & Krug, D. (1992). Responding to feedback after multiple-choice answers: The influence of response confidence. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 45*A*, 649-667.

Sturges, P. T. (1969). Verbal retention as a function of the informativeness and delay of informative feedback. *Journal of Educational Psychology*, 60, 11-14.

Sturges, P. T. (1972). Information delay and retention: Effect of information in feedback and tests. *Journal of Educational Psychology*, 63, 32-43.

Sturges, P. T. (1978). Delay of informative feedback in computer-assisted testing. *Journal of Educational Psychology*, 70, 378-387.

Sullivan, H. J., Schutz, R. E., & Baker, R. L. (1971). Effects of systematic variations in reinforcement contingencies on learner performance. *American Educational Research Journal*, 8, 135-142.

Surber, J. R., & Anderson, R. C. (1975). Delay-retention effect in natural classroom settings. *Journal of Educational Psychology*, 67, 170-173.

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

Tulving, E. (1983). *Elements of episodic memory.* New York: Oxford.

Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology*, 19, 251-265.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, 16, 465-478.

BIOGRAPHICAL STATEMENT

Troy A. Smith is a cognitive scientist with interests in understanding human learning and memory from a theoretical perspective and applying the theoretical principles of cognitive psychology to improve education and training. Troy has 6 years of experience as a professional educator, including 3 years teaching at the associate degree level and 3 years teaching high school mathematics. Prior to becoming a professional educator, he served in the U.S. Navy for 6 years as an electronic warfare technician. Troy has an A.A. degree in electronics for the University of Pheonix, an Honors B.S. degree in Interdisciplinary Studies with a professional focus on technical education and training from The University of Texas at Arlington, and (upon successful completion of this thesis) a M.S. degree in Psychology from The University of Texas at Arlington. His current research interests include computational modeling of false memory, investigating the effectiveness of learning optimization techniques, and developing a model of adaptive expertise.