THE EFFECTS OF IMMEDIATE VERSUS DELAYED FEEDBACK AFTER MULTIPLE-CHOICE QUESTIONS ON SUBSEQUENT EXAM PERFORMANCE

By

NEHA SINHA

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Psychology

Written under the direction of

Dr. Arnold Glass

And approved by

_____

_____

_____

New Brunswick, New Jersey,

October, 2012

ABSTRACT OF THE THESIS

The Effects of Immediate versus Delayed Feedback after Multiple-Choice Questions on Subsequent Exam Performance

By NEHA SINHA

Thesis Director:

Dr. Arnold Glass

This thesis investigates the effects of immediate versus delayed feedback following multiple-choice questions on subsequent performance on multiple-choice and recall questions. In three experiments, students in a college psychology lecture course received immediate or delayed feedback following multiple-choice questions on an initial unit exam which was followed up with exam(s) including both multiple-choice and short-answer questions. In the first experiment, the kind of feedback did not affect performance on the same multiple-choice questions when they were repeated on the final. In the second experiment, two subsequent follow-up exams included first a short-answer version of the multiple-choice question and then the same multiple-choice question. Performance on the short-answer questions was better following delayed feedback than following immediate feedback. However, the kind of feedback had no effect on the performance of the repeated multiple-choice questions. Also, the interval between the

initial exam and the follow-up exam had no effect on performance. The third experiment examined whether delayed feedback increased confidence more than immediate feedback and whether the increase in confidence mediated the improved performance on subsequent short-answer questions. The delayed feedback had no effect on confidence for the subsequent short-answer and multiple-choice responses. Together, these results demonstrate that delayed feedback improves performance on the short-answer questions by increasing the subsequent generation of the correct response but does not influence recognition of it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Taking a test generally improves retention of the material tested—a result commonly called the *testing effect* (Roediger & Karpicke, 2006). Furthermore, providing the correct answer as feedback after the student's response further improves performance (Butler, Karpicke, & Roediger, 2007; Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008). The correct answer to a question may be presented immediately after the question is answered or after a subsequent delay. Research comparing the effects of immediate versus delayed feedback on learning began with Pressley (1926) and was stimulated by Skinner (1954). However, when Kulik and Kulik (1988) performed a comprehensive review of experimental studies of the effect of feedback their literature search did not find a single experimental report that had been published in the 1980's. Clearly, interest in the effect of feedback on learning had waned at the time. The reason, as we shall see below, was that this early research had yielded inconsistent results.

Recently, Butler, et al. (2007), Butler, et al. (2008), and Butler and Roediger (2008) have again taken up research on the effect of feedback. Their studies were concerned with the retention of the study material, as measured by performance on a follow-up exam. To this end Butler and his colleagues compared the effect of immediate feedback with the effect of delayed feedback during an initial multiple-choice exam on performance on a subsequent short-answer exam. Presenting the answer to a question immediately after a student has responded is called immediate feedback. Withholding the answer to the question until the student has answered several additional questions is called delayed feedback. Two laboratory studies (Butler, et al., 2007; Butler & Roediger,

2008) found that delayed feedback on the initial multiple-choice test led to better performance on the follow-up cued recall test than immediate feedback. Butler and his colleagues pointed out that distributed repetition of study material produced higher levels of retention than massed repetition (Schmidt & Bjork, 1992). They suggested that delayed feedback was more effective for this reason.

The findings of Butler and his colleagues were entirely consistent with the Kulik and Kulik (1988) review of previous research. Kulik and Kulik reported that **laboratory** studies comparing delayed with immediate feedback found that delayed feedback produced better performance on the follow-up exam. However, they reported the opposite result for **classroom** studies of feedback. When the experiment was conducted with tests and materials that were assigned for a course, delayed feedback did not produce better performance on the follow-up exam. Either immediate feedback during a multiple-choice exam produced better performance on a follow-up exam than delayed feedback or else there was no difference. To explain why eight previous classroom studies had not found better subsequent performance following delayed feedback, Butler et al. (2007) made the post hoc suggestion that in a classroom the students might not have fully processed the delayed feedback. Butler provided no supporting evidence for this idea. Nor did he explain why, if students are routinely so uninterested in their grades as to not process delayed feedback that could improve subsequent exam performance, hence their grades, in eight different studies, they nevertheless assiduously processed it in a laboratory while performing a task that had no benefit to the students themselves beyond

receiving payment or credit for their participation regardless of their level of performance.

Therefore, the purpose of this study was to further compare the effects of immediate versus delayed feedback following multiple-choice questions on subsequent exam performance. Experiment 1 compared their effects on subsequent multiple-choice questions in the classroom. Experiments 2 and 3 compared their effects on both subsequent multiple-choice questions and subsequent short-answer questions in the classroom.

2 Experiment 1

The purpose of Experiment 1 was to perform an experiment similar to those of Butler and his colleagues in the classroom, but under conditions in which the students would attend closely to the delayed feedback. An experimental design to test the effects of delayed versus immediate feedback was embedded in an upper level college psychology lecture course. In the course-embedded experimental paradigm, the study and test items in an academic course are also the materials used in the experiment. In a multi-section course different items are presented in different conditions to different sections, thus making a within-subject, within-item design possible.

The experiment was similar to those performed by Butler and Roediger except that the follow-up exam was a multiple-choice exam rather than a cued recall exam. Two 36-question unit exams during the course were administered as clicker exams. Each multiple-choice question was presented on a Power Point slide and the students responded using personal response devices (clickers). The first 18 questions presented were in the delayed feedback condition. After time expired for a response the next question was presented. The final 18 questions were presented in the immediate feedback condition. After time expired for a response, a large green check mark appeared next to the correct answer, and the students had 10 seconds to study the answer. Immediately after the final question each of the first 18 questions was again presented, one at a time, this time with the correct answer indicated. It took about 20 minutes to answer 18 questions, so the delayed feedback followed the corresponding questions by about 20 minutes. In comparison Butler, et al. (2007) and Butler and Roediger (2008)

found that delayed feedback was more effective than immediate feedback over feedback-delays of 10 minutes through one day.

The two unit exams were presented during weeks 9 and 14 of the course and the final exam was presented during week 16 of the course. The students were instructed on the syllabus and in class that the final exam would contain questions identical or very similar to all the questions on the two unit exams. They were told to use the feedback on the unit exams as study opportunities to improve their performance on the final exam. Students were instructed that in both the immediate and delayed feedback conditions they should imagine that the correct response as their response because if they did so, that would improve their performance when the same or a similar question appeared on the final exam. Each unit exam counted for one-sixth of the final grade and the final exam counted for one-half of the final grade, so the students were highly motivated to process both the immediate and delayed feedback. (Another unit exam, not included in the experiment, counted for the remainder of the final grade.)

Another purpose of the experiment was to determine whether providing immediate feedback on a unit exam would affect performance on that exam. Glass (2009) had previously found that performance on a clicker exam without feedback was virtually identical to performance on a paper and pencil version of the exam. However, it seemed possible that immediate feedback might disrupt performance. Would knowing that their most recent response was incorrect upset a student or otherwise influence her criterion so that she was less likely to answer a subsequent question correctly? Increased anxiety might increase the likelihood that the next question was misread or make retrieval

of the correct answer more difficult or decrease confidence in the response, so that the student was more likely to guess. So, Experiment 1 was designed to determine whether providing immediate feedback after every question reduced performance on an exam.

2.1 Method

The experimental design was embedded in two sections of a summer session psychology course on memory offered at a state university.

Participants

A total of 377 students participated in the experiment. Each student was enrolled in one of two sections of the same course. The students answered three voluntary demographic questions. There were 137 males and 191 females. The students also self-identified themselves as 11 African-American, 86 Asian, 168 Caucasian, 11 Latino, 20 mixed race, and 32 other. The students also self-identified themselves as 18 years old or younger (8 students), 19 – 24 (312 students), 25 -36 (7 students), and 37 – 99 years old (1 student). As can be seen by summing across responses for each category, no question was answered by all 377 students.

Experimental Materials

The materials consisted of 68 question sets containing the Pre-Class homework (H), In-Class (C), and Exam (E) multiple-choice questions such that H-questions were presented before class, the C-questions were presented in class, and the E-questions were presented on a unit exam and again on the final exam. It was always the case that a

single proposition logically entailed the answer to all three members of the question set. An example question set is presented in the Appendix.

Procedure

The semester was 16 weeks long. Fourteen weeks of instruction were followed by a two-week reading and final examination period. There was a reading assignment in the textbook that was relevant to each lecture. All of the reading assignments for the entire semester were listed in the course syllabus, which was posted on the course website before the semester began. An online quiz corresponding to the next class was made available at the end of each class. The online quiz always consisted of the Pre-Class (H) questions and tested students on the reading assignment in the textbook for the next class. The quiz was graded as soon as it was completed and students received immediate feedback consisting of the correct answers and an explanation of the correct answer that included a quotation from the textbook providing the answer along with its page and paragraph location in the textbook. Students were aware from the syllabus and instructions in class that if their online-quiz-score was greater than their exam score, their exam score would be increased.

During the following lecture, at the appropriate moments, each of the In-Class (C) questions was presented as clicker questions. The question was presented as a Power Point slide and the students responded by using personal response devices (clickers). The correct answer was presented immediately after the responses were made. Students were

aware from the syllabus and instructions in class that if their in-class-quiz-score was greater than their exam score, their exam score would be increased.

Thirty-six of the question sets covered material presented on weeks 5 through 9 of the course and the 36 E-questions comprised a unit exam presented during week 9. The remaining 32 of the question sets covered material presented on weeks 10 through 14 of the course and the 32 E-questions comprised a unit exam presented during week 14. All 68 E-questions again appeared on the final exam at the end of week 16.

Each unit exam consisted of the 36 or 32 Exam (E) questions. During the exam, each question was presented on a Power Point slide, one at a time, for either 55 seconds or 70 seconds, and students responded with clickers. During previous semesters, the time required for 90% of the class to answer each question was recorded. Questions that were presented for 55 seconds received responses from 90% of the class in no more than 45 seconds. These were questions for which each of the five alternatives were no more than each four words or less. Questions that were presented for 70 seconds received responses from 90% of the class in no more than 60 seconds. These were questions for which at least some of the alternatives comprised a phrase of five or more words.

The first half of the exam was the delayed feedback condition. For the first half of the exam, after time ran out, the next question was presented without feedback as to the correct answer. The second half of each exam was the immediate feedback condition. In the second half of the exam, after time ran out, a large green check mark appeared next to the correct answer before the next question was presented. The students had 10

seconds per question to absorb the feedback. After the last question, each question from the first half of the exam was again presented briefly (30 seconds) with the correct answer marked. Note that students were given slightly more time while presenting delayed feedback. This was because, in the immediate feedback condition, having just answered the question they only needed time to focus on the feedback being given whereas while getting delayed feedback, they would have to go over the entire question once again before they could process the feedback.

The students knew that questions very similar or identical to the questions on the unit exam would appear on the final exam. So they were highly motivated to attend to the feedback. Furthermore, before the immediate condition the students were told to pay close attention to each correct answer when it was presented and to imagine that it was the answer they gave. They were told in order to increase their chance of responding correctly to the same question on the final exam they should put all of their effort into learning the correct response rather than trying to remember whatever answer they gave. The same instructions were given before presenting the feedback for questions from the delayed condition.

Hence, 18 of the 36 questions on one unit exam were presented with immediate feedback and 18 of the 36 questions were presented with delayed feedback. On the other unit exam, 16 of the 32 questions were presented with immediate feedback and 16 of the 32 questions were presented with delayed feedback. The questions presented with immediate feedback to one section of the course were presented with delayed feedback to

the other section of the course. Within each condition, the questions were presented in the order in which their topics had been covered during the class.

All 68 E-questions were again presented on the final exam. The final was again a clicker exam. Students saw each question on a Power Point slide for the same amount of time that it had been presented on the unit exam.

2.2 Results

A criterion of $p < .05$ was adopted for all analyses in this report.

An analysis was performed on the corresponding homework and clicker questions presented before the questions on unit exam that appeared in each condition. On the homework, percent correct was 72% (C.I. = 67, 76) when the questions preceded exam questions in the immediate feedback condition and 72% (C.I. = 67, 75) when the following exam questions were in the delayed feedback condition. On the clicker questions, percent correct was 76% (C.I. = 73, 81) when the questions preceded the immediate feedback condition and 75% (C.I. = 71, 79) when they preceded the delayed feedback condition. A 2 x 2 analysis of variance in which items was the random factor was performed on the effects of Question type (Homework (H) versus Clicker (C)) and Feedback subset (following exam questions were in delayed feedback condition versus following exam questions were in the immediate feedback condition. The effects of Feedback, $F(1,132) = .46$, $p = .5$, Question type, $F(1,132) = 2.2$, $p = .14$ and the interactions between Feedback and Question type, $F(1,132) = .85$, $p = .36$ were not significant.

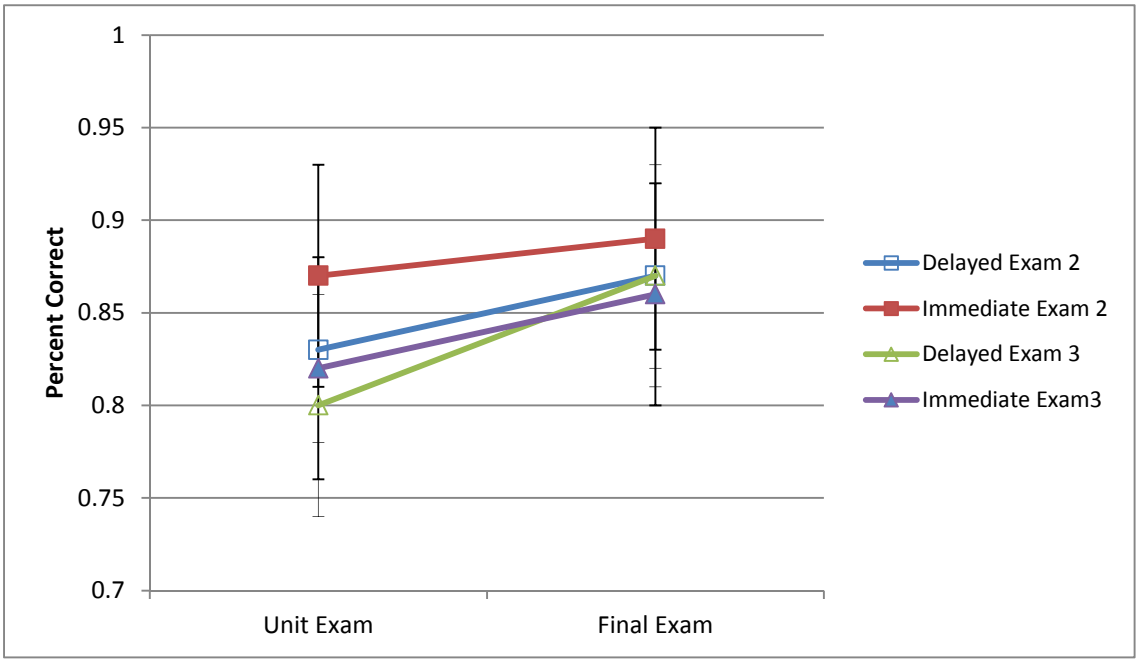

Figure 1: The effect of delayed versus immediate feedback during a unit exam on the percent correct on the same multiple-choice question on the final exam

Percent correct on the unit exams and the final exam is shown in Figure 1. On the unit exams, percent correct was 84% (C.I. = 83, 85) for the immediate feedback condition and 82% (C.I. = 80, 83) for the delayed feedback condition. An analysis of variance in which subjects was the random factor was performed on the effects of feedback on the unit exam (delayed versus immediate). The effect of feedback, $F(1, 376) = 27.2$ was significant. The same analysis was performed using items as the random factor. In this analysis, the effect of feedback, $F(1,67) = 3.75$, $p = .06$ was **not** significant. Consequently, when the results of the subject and item analyses were used to compute the min-F'statistic, the effect of feedback, $F'(1, 86) = 3.3$ was **not** significant.

On the final exam, percent correct was 87% (C.I. = 86, 88) for the questions that had previously appeared in the immediate feedback condition and 87% (C.I. = 86, 88) for the questions that had previously appeared in the delayed feedback condition. The effect of feedback over subjects, F(1,376)= .278, p = .6 and items, F(1, 67) = 2, p = .16 was **not** significant. The value of min F, F'(1, 401) = .47, p= .6 was, not significant.

2.3 Discussion

Percent correct did not differ between the homework and classroom questions that preceded the corresponding exam questions that received delayed feedback and the corresponding exam questions that received immediate feedback. So before the exam the students did not have different levels of knowledge of the fact statements answering the questions for which they received delayed feedback versus those for they received immediate feedback on the exam.

The Effect of Immediate Feedback on Exam Performance: Immediate feedback increased performance on the unit exams from 82% to 84%. This result was significant when subjects was treated as a random effect. What this means is that if different subjects were given **exactly the same questions** then it is likely that immediate feedback would again result in better performance. However, Clark (1973) pointed out that often the investigators want to draw a broader conclusion. The investigator wants to conclude that the effect (e.g. of feedback) is more general and would produce the same results if different students took different exams. However, this conclusion cannot be drawn without the correct statistical test being done to support it. One can no more claim generality from the fact that an investigator has examined the effect on a large and

diverse number of items, without actually computing the statistical test of the generality of the effect, than one can claim generality over subjects, regardless of the number of subjects, without actually computing the statistical test of the generality of that effect. One reason for this, as Clark pointed out, is that a large and reliable effect over subjects may be obtained from a small number of the items tested. Collapsing over items and failing to test the effect independently over them completely obliterates the evidence of this.

To test for generalization over items, Clark (1973) proposed doing analyses in which items were treated as the random effect. He then proposed combining the results of the subject and item analyses to compute a min-F statistic that would permit generalization over both subjects and items. The min-F is an approximation of the true F statistic. The sole reason for offering this alternative is that the most commonly used statistical computing packages did not then (and do not now) provide a simple method for computing the true F statistic for two random effects and the min-F is easy to compute by hand. Subsequent computer modeling demonstrated that the alpha value indicated by the min-F statistic is inaccurate and the true alpha value is much smaller, .i.e. $p = .01$ is actually .001 or less. This means that the test is highly conservative and significant min-F indicates that the result will replicate over both subjects and items.

In the absence of a confirming significant effect over items, a possible implication of the significant effect difference between immediate versus delayed feedback when subjects was the random effect is merely that the students in one section happened to know the answers to a few of the questions better than the students in the

other section and the students who knew the answers better happened to receive the unit exam versions of the questions in the immediate feedback condition. Strong support for this implication is provided by the fact that the students had previously performed better on the classroom versions of these questions though the presentation of the classroom versions of the questions was identical for both sections.

So, when the failure to find an effect over items is considered, one cannot reject the null hypothesis and must conclude that there was no effect of delayed versus immediate feedback on performance on the unit exams.

The Effect of Delayed Feedback on Subsequent Performance: More importantly, there was no subsequent effect of delayed versus immediate feedback on the unit exam on performance on the final exam. The superior effect of delayed feedback found by Butler, et al. (2007) and Butler and Roediger (2008) in the laboratory was not replicated in the classroom in this experiment.

Furthermore, a detailed examination of the responses to questions that students got wrong on both the unit and final exams confirmed that there was no difference as the result of the type of feedback. As shown on the top row of Table 2, on the unit exam, a student averaged 5 errors in the delayed feedback condition and 5 errors in the immediate feedback condition. On the final exam, for those questions that had been in the delayed feedback condition that the students again got wrong, they selected their previous wrong answer 23% of the time. For those questions that had been in the immediate feedback condition that the students again got wrong, they selected their previous wrong answer 20% of the time. Since these were five-alternative multiple-choice questions, the

implication of selecting the same response about 20% of the time is that the students remembered neither their previous answer nor the correct answer (from the feedback) on the unit exam and hence guessed among the five alternatives.

The failure to find an advantage of delayed versus immediate feedback in a classroom experimental study is completely consistent with the results of the previous classroom experimental studies reviewed by Kulik and Kulik (1988). Finally, given the motivation of the students and the methodology that directed their attention to the feedback, the failure to find an effect of delayed feedback can not be attributed to a failure to attend to it.

We will consider two possible explanations for the difference in results obtained here and by Butler and his colleagues.

First, even though the design had sufficient power to detect a difference between 82% and 84%, a ceiling effect on student performance eliminated differences resulting from feedback when mean performance was at 87%. However, half of the nearly four hundred students participating performed at worse than 87%, providing a range of scores over which an effect of delayed feedback on subsequent performance could be observed if it actually existed.

Furthermore, there were two notable differences between the experiments of Butler and his colleagues and this one. First, the Butler, et al. (2007) and Butler and Roediger (2008) laboratory studies used a short-answer version of the multiple-choice study question to test retention rather than repeating the multiple-choice question. Second, the laboratory studies made use of a shorter retention interval between the initial

exam and the follow up exam. In the three experiments in these two laboratory studies, the retention interval was either one day or one week. In comparison, in this experiment, the retention interval was either 2 weeks or 7 weeks. In Experiment 1, as shown in Figure 1, though not significant, there was a suggestive relationship between retention interval and the effect of delayed feedback on the final exam. When the interval between the unit exam and the final exam was 7 weeks, percent correct on the final following delayed feedback was 87.2% (C.I. = 86.1, 88.3) and percent correct on the final following immediate feedback was 88.7% (C.I. = 87.7, 89.7), which is in the opposite direction from the results of Butler and his colleagues. When the interval between the unit exam and the final exam was 2 weeks, percent correct on the final following delayed feedback was 87.0% (C.I. = 85.7, 88.2)and percent correct on the final following immediate feedback was 86.4% (C.I. = 85.1, 87.4), which is in the same direction as the results of Butler and his colleagues. As these confidence intervals indicate, none of these differences approached significance, nor did the interaction between the effects of retention interval and feedback. However, the fact that effect of delayed feedback was only in the same direction as the laboratory experiments at the shorter retention interval, and that even this interval was longer than those used in the laboratory experiments, raises the possibility that at intervals identical to those used in the laboratory studies the results would be the same.

The failure to find a differential effect of immediate versus delayed feedback on subsequent performance was consistent with the results of previous experimental studies in the classroom reviewed by Kulik and Kulik (1988). Recall that they found that

immediate feedback produced better subsequent performance or there was no difference. In this experiment, the questions on the follow-up exam were multiple-choice questions rather than short-answer questions to avoid the burden of scoring 72 short-answer questions for 377 students. Since other experimental studies in the classroom would have faced the same burden, we decided to check the kinds of questions asked on the follow-up exams in the studies reviewed by Kulik and Kulik, which produced the same result.

Kulik and Kulik's review included seven reports of studies comparing feedback after each item with feedback after the test. We reviewed four of the reports (Angell, 1949; Pressey, 1926; Saunderson, 1974; Sullivan, Schultz, & Baker, 1971). (Two of the remaining reports had appeared in limited circulation publications and were not available in our library and one had not been published at all.) We found that in three of the studies (Angell, 1949; Pressey, 1926; Sullivan, et al., 1971), the follow-up exam was a multiple-choice exam and the fourth study (Saunderson, 1974) did not describe the kind of exam. So the distinction between laboratory and classroom experimental studies in Kulik and Kulik's (1988) review is at least partly, and may be entirely, confounded with whether the follow-up exam was a multiple-choice exam or a short-answer exam. Furthermore, an additional classroom study not cited by Kulik and Kulik (White, 1968) found that immediate feedback produced better performance on a follow-up exam than delayed feedback did, which is consistent with the results of both Kulik and Kulik's review and Experiment 1.

3 Experiment 2

The purpose of Experiment 2 was to investigate the effects of two differences between the design of Experiment 1 and the experiments of Butler and his colleagues. To assess the effect of retention interval, the initial multiple-choice exam was followed by a short-term retention test one week later and a long-term retention test three weeks later. One week was equal to the largest retention interval used by Butler and his colleagues. Three weeks was longer than the shortest retention interval in Experiment 1. Therefore, if the longer retention interval were the reason that delayed feedback did not result in better performance on the follow-up exam in Experiment 1, then delayed feedback should improve performance on the short-term retention test but not on the long-term retention test.

A summer session psychology course provided the opportunity for the experiment. During the six weeks of the summer session course it was possible to construct a course-embedded design such that the intervals between the initial and final exams were one and three weeks, respectively. However, the summer session course had a much smaller enrollment, hence a smaller subject sample, than the regular academic year course in which Experiment 1 was embedded.

To assess the effect of question type, the retention tests included both short-answer and multiple-choice questions. The short-term retention exam contained one-half and the long-term retention exam contained the other half of the questions that appeared on the initial multiple-choice test. Each retention exam consisted of two parts. The first half of the exam was a short-answer test that contained short-answer versions of the

initial multiple-choice questions. The second half of the exam was a multiple-choice test that repeated the multiple-choice questions. Hence, first the short-answer version and then the multiple-choice version of a question appeared on the exam. If the difference in type of question was the reason that delayed feedback did not result in better performance on the follow-up exam in Experiment 1, then there should be better performance on the short-answer questions following delayed feedback but not on the multiple-choice questions following delayed feedback. If the difference in the longer retention interval was the reason that delayed feedback did not result in better performance on the follow-up exam in experiment 1, then there should be better performance on both the short-answer and multiple-choice questions following delayed feedback on the short-term retention test but not the long-term retention test.

3.1 Method

The experimental design was embedded in two sections of a summer session psychology course on memory offered at a state university.

Participants

A total of 35 students participated in the experiment. There were 13 males and 22 females. Students self-reported their ethnicity as follows, 1 African, 12 Asian, 18 Caucasian, 1 Latino, 1 Mixed, and 2 other. The students also self-identified themselves as 18 years old or younger (1 student), 19 – 24 (29 students) and 25 -36 (5 students).

Experimental Materials

The materials consisted of 36 question sets containing the Pre-Class (H), In-Class (C), and Exam (E) multiple-choice questions such that H-questions were presented before class, the C-questions were presented in class, and the E-questions were presented on the initial exam and again on either the short-term or long-term retention exam. In addition, for each E-question a corresponding Es-question short-answer question was constructed. Each short-answer question was presented on either a short-term or long-term retention exam.

It was always the case that a single proposition logically entailed the answer to all four members of the set. An example question set used in the experiment is presented in the Appendix.

All 36 multiple-choice questions had been asked during the same course during previous semesters, and the percent correct responses for each question had been tabulated. The 36 question sets were partitioned into four subsets, A, B, C, D, such that the mean percent correct for the E-questions in each subset was equal. Also, the questions were partitioned such that percent correct for the corresponding C and H questions was also as equal as possible.

During the previous semesters, all the questions in the 36 sets consistently had positive biserial correlations between percent correct on the question and percent correct on the entire quiz or exam on which the question appeared. A positive biserial correlation indicates that a student who knew more answers overall was more likely to answer that question correctly.

Procedure

Each summer session section was 6 weeks long. There was a reading assignment in the textbook that was relevant to each lecture. All of the reading assignments for the entire semester were listed in the course syllabus, which was posted on the course website before the semester began. An online quiz corresponding to the next class was made available at the end of each class. The online quiz always consisted of the Pre-Class (H) questions and tested students on the reading assignment in the textbook for the next class. The quiz was graded as soon as it was completed and students received immediate feedback consisting of the correct answers and an explanation of the correct answer that included a quotation from the textbook providing the answer along with its page and paragraph location in the textbook. Students were aware from the syllabus and instructions in class that if their online-quiz-score was greater than their exam score, their exam score would be increased.

During the following lecture, at the appropriate moments, each of the In-Class (C) questions was presented as clicker questions. The question was presented as a Power Point slide and the students responded by using personal response devices (clickers). The correct answer was presented immediately after the responses were made.

The students knew from the syllabus and repeated announcements in class when the initial exam and two follow-up exams would occur and what the initial exam would cover. They knew that all of the material on the initial exam would be repeated on a

follow-up exam. They knew that their performance on the three exams would be included in the computation of their grade for the course.

At the end of lecture 3, the initial exam was presented, consisting of the 36 Exam (E) questions. This covered about one-third of material in the course, which was similar to what was covered by one unit exam in Experiment 1. During the exam, each question was presented on a Power Point slide, one at a time, and students responded with clickers, as described for Experiment 1. The first half of the exam was the delayed feedback condition. For the first half of the exam, after time ran out, the next question was presented without feedback as to the correct answer. The second half of each exam was the immediate feedback condition. In the second half of the exam, after time ran out, a large green check mark appeared next to the correct answer before the next question was presented. The students had 10 seconds per question to absorb the feedback. After the last question, each question from the first half of the exam was again presented briefly (30 seconds) with the correct answer marked. As in experiment 1, the students were given slightly more time while presenting delayed feedback because they would have to go over the entire question once again whereas in the immediate feedback condition, having just answered the question they only needed time to focus on the feedback being given. (Recall that the students knew that they would be tested on this material again and that their subsequent performance would influence their course grade. So they were highly motivated to attend to the feedback.)

Hence, 18 of the 36 questions were presented with immediate feedback and 18 of the 36 questions were presented with delayed feedback. All the questions on the exam were about the material covered in the previous three lectures.

Table 1: The Design of Experiment 2

Retention Duration		Short		Long	
Feedback		Immediate	Delayed	Immediate	Delayed
Section 1	Group 1	A	C	B	D
	Group 2	B	D	A	C
Section 2	Group1	C	A	D	B
	Group 2	D	B	C	A

Table 1 shows the design of the experiment. As mentioned above, the questions had been partitioned into four subsets of nine questions each. On the initial exam, one section of the course saw the questions in subsets A and B in the delayed feedback condition and the questions in subsets C and D in the immediate feedback condition. The reverse was true for the other section. So, the questions that were presented to one section with immediate feedback were presented to the other section with delayed feedback and vice versa. Otherwise, the questions that the two sections answered were identical. Within each condition, the questions were presented in the order in which their topics had been covered during the class.

Each section was divided in two subgroups. Each subgroup within a section received a different short-term retention exam one week after the initial exam and a different long-term retention exam three weeks after the initial exam. Half of the questions on each follow-up exam had initially appeared with delayed feedback and half of the questions had initially appeared with immediate feedback. As shown in Table 1, the follow-up exam that was used as the short-term retention exam for one subgroup of the section was used as the long-term retention exam for the other subgroup of the section.

On each follow-up exam first 18 short-answer questions were presented without feedback. Then, the corresponding 18 multiple-choice questions from the initial exam were again presented. As can be seen in Table 1, each of the four groups of students got a different one of the four subsets of questions in each of the four experimental conditions.

3.2 Results

An analysis was performed on the corresponding homework and clicker questions presented before the questions on unit exam that appeared in each condition. On the homework, percent correct was 62% (C.I. = 53, 69) when the questions preceded exam questions in the immediate feedback condition and 64% (C.I. = 57, 71) when the following exam questions were in the delayed feedback condition. On the clicker questions, percent correct was 69% (C.I. = 60, 77) when the questions preceded the immediate feedback condition and 71% (C.I. = 63, 79) when they preceded the delayed

feedback condition. A 2 x 2 analysis of variance in which items was the random factor was performed on the effects of Question type (Homework (H) versus Clicker (C)) and Feedback subset (following exam questions were in delayed feedback condition versus following exam questions were in the immediate feedback condition. The effects of Feedback, $F(1,33) = .22$, $p = .64$, Question type, $F(1,33) = 2.7$, $p = .11$ and the interactions between Feedback and Question type, $F(1, 33) = .98$, $p = .33$, were not significant.

Next, the result of the initial exam was analyzed. The difference on percent correct between the delayed feedback condition, 70.8 (C.I. = 67, 75), and the immediate feedback condition, 71 (C.I. = 68, 75), was not significant, $t(34) = .259$, $p = .8$.

Second, the results of the two follow-up exams were analyzed. The results are shown in Figure 2. A 2 x 2 x 2 analysis of variance in which subjects was the random factor was performed on the effects of Feedback on the initial exam (delayed versus immediate), the Retention duration (short versus long), and the Question type (short-answer versus multiple choice). The effect of feedback, $F(1,34)= 11.7$, was significant. Also, the interaction between feedback and question type, $F(1,34)= 27.9$, was significant. However, as can be seen by comparing the top and bottom panels of Figure 2, the effect of retention duration was not significant, $F(1,34) = .237$, $p = .63$ and did not participate in any significant interactions. The same analysis was performed using items as the random factor. In this analysis also, the effect of feedback, $F(1,35)= 6$, and the interaction between feedback and question type, $F(1,35) = 16.1$ was significant. The results of the subject and item analyses were used to compute the F'statistic. Again, feedback,

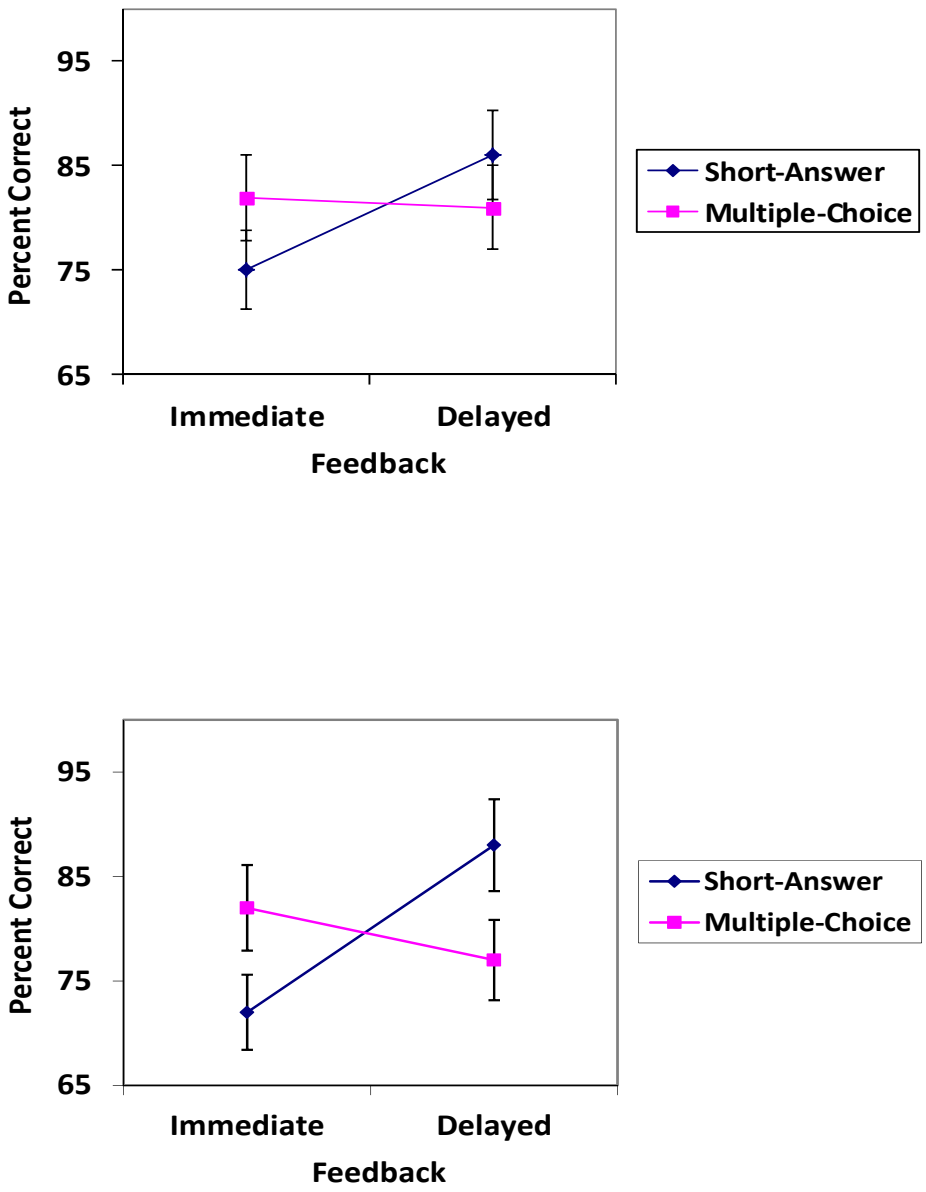

Figure 2: The effect of immediate and delayed feedback on multiple-choice questions on percent correct for subsequent short-answer and multiple-choice questions after short-term (top panel) and long-term (bottom panel) retention.

F'(1,63) = 4, and the interaction between feedback and question type, *F'*(1,67) = 10.2 were significant.

Orthogonal contrasts tested the locus of the interaction. Percent correct on the short-answer questions was significantly greater following delayed feedback than following immediate feedback, *t*(136) = 4.38, d=1.01. Percent correct on the multiple-choice questions was **not** significantly greater following immediate feedback than following delayed feedback, *t*(136) = .93, *p* = .35.

3.3 Discussion

As was the case for Experiment 1, percent correct did not differ between the homework and classroom questions that preceded the corresponding exam questions that received delayed feedback and the corresponding exam questions that received immediate feedback. So before the exam there was no evidence that the students had different levels of knowledge of the fact statements answering the questions for which they received delayed feedback versus those for they received immediate feedback on the exam.

Unlike Experiment 1, there was only a tiny difference in unit exam performance for immediate versus delayed feedback, which was not significant when subjects was treated as the random effect in the analysis. However, since the effect of feedback in Experiment 1 was not significant over items, the failure to replicate over different items in Experiment 2 is not inconsistent.

On both follow-up exams, students did better on the short-answer questions for which delayed feedback had been provided on the initial exam. In contrast, students did not do better on the multiple-choice questions for which delayed feedback had been provided on the initial exam. In fact, percent correct was higher for those multiple-choice questions that had received immediate feedback, but this difference was not significant. Both of these results are consistent with previous findings. The superior performance on the short-answer questions following delayed feedback on the initial multiple-choice exam is consistent with the finding of Butler, et al. (2007) and Butler and Roediger (2008) and the laboratory studies reviewed by Kulik and Kulik (1988). Furthermore, the effect size is 1.01 which is greater than the largest effect size of .56 observed in the two laboratory studies conducted by Butler and his colleagues. The inferior performance on the multiple-choice questions following delayed feedback on the initial multiple-choice exam is consistent with the finding of Experiment 1 and the classroom studies reviewed by Kulik and Kulik (1988). However, this is the first experiment, either in the laboratory or in the classroom, in which the follow-up exam included both short-answer and multiple-choice questions. Hence, this experiment has produced the important, novel result that delayed feedback only improves performance on a short-answer question on a subsequent exam but not on a multiple-choice question. This effect was obtained in a counter-balanced, within-subject, within-item design and was found significant for both subjects and items. This is a very conservative statistical test. So the result is highly robust.

Furthermore, a detailed examination of the responses to questions that students got wrong on both the unit and final exams confirmed that there was no difference in performance on the multiple-choice questions as the result of the type of feedback. As shown on the middle row of Table 2, on the unit exam, a student averaged 5 errors in the delayed feedback condition and 5 errors in the immediate feedback condition. On the final exam, for those questions that had been in the delayed feedback condition that the students again got wrong, they selected their previous wrong answer 19% of the time. For those questions that had been in the immediate feedback condition that the students again got wrong, they selected their previous wrong answer 15% of the time. Since these were five-alternative multiple-choice questions, the implication of selecting the same wrong response less than 20% of the time when they again got the question wrong is that the students did not remember the correct answer from the feedback on the unit exam and sometimes did not remember, hence avoid, their previous wrong answer.

We will consider three possible explanations for the difference in results obtained here versus those of Butler and his colleagues.

The first explanation is that even though the design had sufficient power to detect an increase from the 73.5% percent correct observed for the short-answer questions following immediate feedback to a higher percent correct following delayed feedback, it did not have sufficient power to detect an increase from the 82% percent correct observed for the multiple-choice questions following immediate feedback to a higher percent correct following delayed feedback. If this ceiling effect explanation was correct, then percent correct for multiple-choice questions would be greater than percent correct for

short-answer questions following immediate feedback but there would be no difference between percent correct for short-answer questions and percent correct for multiple-choice questions following delayed feedback because of the ceiling effect on percent correct which limited the increase for the multiple-choice questions. However, as shown in Figure 2, while percent correct for multiple-choice questions was significantly greater than percent correct for short-answer questions following immediate feedback, $t(34) = 3.7$, the reverse was the case and percent correct for short-answer questions was significantly greater than percent correct for multiple-choice questions following delayed feedback, $t(34) = 2.7$. On the basis of this observed significant difference, the hypothesis that the interaction between feedback and question-type resulted from a ceiling effect can be rejected.

The second explanation is that delayed feedback increases confidence in correct responses more than immediate feedback does and increased confidence increased probability of reporting a generated correct response to a short-answer question than selecting the correct response to a multiple-choice. Butler et al. (2008) found that feedback improved retention because feedback made subjects more confident of low-confidence correct responses. Increased confidence in correct responses increased the probability of a correct response, which ultimately increased percent correct. Two weaknesses in this explanation are first, there is no apparent reason that delayed feedback should increase confidence more than immediate feedback and second, there is no apparent reason that increased confidence should only affect the reporting of generated responses and not the selection of presented reasons. Nevertheless, since feedback has

been shown to influence confidence, assessing the effect of delayed versus immediate feedback on confidence is necessary to determine whether confidence mediated the effect of delayed feedback on subsequent short-answer questions.

The third explanation is that delayed feedback increases the likelihood of generating the correct response but does not influence the probability of recognizing it. Consequently, performance is improved for subsequent short-answer questions, which require generation of the correct answer, but is not improved for subsequent repetition of the multiple-choice questions, which merely require recognition of the correct answers. This explanation exists within the framework of the generate-and-recognize model of recall (Higham & Tam, 2005). This hypothesis describes recall as a two-step process in which the target response is first generated and then recognized. The generate-and-recognize hypothesis is derived from a well supported dual-system description of mammalian memory. Within the framework of the dual-system description, mammalian memory consists of a medial temporal instrumental system, including the hippocampus, and a medial frontal habit system, including the striatum (Packard & McGaugh, 1996; Yin & Knowlton, 2006). The instrumental system recognizes spatial patterns and the habit system generates them. Yin and Knowlton (2006) point out that distribution testing (in animals) increases learning and retention much more in the habit system than in the instrumental system. Extending this result to specifically human memory implies that delaying feedback should increase the likelihood of generating the correct response through an effect on the habit system

4 Experiment 3

The purpose of Experiment 3 was to investigate whether delayed feedback increases confidence more than immediate feedback and whether the increase in confidence mediates the improved performance on subsequent short-answer questions. To assess the effect of feedback (delayed versus immediate) on confidence, the initial multiple-choice exam was followed by final that included both short-answer and multiple-choice questions. After each question, the students were asked to categorize the confidence they would assign to their response as low, medium or high. Unlike Experiment 2, Experiment 3 was embedded in a course during the normal academic year, so the subject sample was much larger.

If the results obtained in Experiment 2 were robust and the failure to find an effect of delayed feedback on subsequent multiple-choice questions was **not** the result of a ceiling effect exacerbated by a small sample size, then they should be replicated and delayed feedback should have a positive effect on short-answer responses, and no effect on multiple-choice responses. If the reason that delayed feedback resulted in better performance on subsequent short-answer versions of the questions than immediate feedback was because it made students more confident, then the average confidence rating for the short-answer questions should be greater following delayed feedback than following immediate feedback but there should be no difference in average confidence ratings for the multiple-choice questions. However, if the effect of feedback is unrelated to the effect of delayed presentation then there should be no difference in confidence for

the short-answer questions following delayed feedback versus those following immediate feedback.

4.1 Method

The experimental design was embedded in two sections of a fall session psychology course on memory offered at a state university.

Participants

A total of 385 students participated in the experiment. Each student was enrolled in one of two sections of the same course. The students answered three voluntary demographic questions. There were 147 males and 199 females. The students also self-identified themselves as 10 African-American, 103 Asian, 162 Caucasian, 27 Latino, 22 mixed race, and 21 other. The students also self-identified themselves as 18 years old or younger (4 students), 19 – 24 (332 students), and 25 -36 (9 students). As can be seen by summing across responses for each category, no question was answered by all students.

Experimental Materials

The materials consisted of 18 question sets containing the Pre-Class (H), In-Class (C), and Exam (E) multiple-choice questions such that H-questions were presented before class, the C-questions were presented in class, and the E-questions were presented on the third unit exam and again on the final. In addition, for each E-question a corresponding Es-question short-answer question was constructed. Each short-answer question was presented on the final exam. It was always the case that a single proposition logically

entailed the answer to all four members of the set. An example question set used in the experiment is presented in the Appendix.

Procedure

The semester was 16 weeks long. Fourteen weeks of instruction were followed by a two week reading and final examination period. There was a reading assignment in the textbook that was relevant to each lecture. All of the reading assignments for the entire semester were listed in the course syllabus, which was posted on the course website before the semester began. An online quiz corresponding to the next class was made available at the end of each class. The online quiz always consisted of the Pre-Class (H) questions and tested students on the reading assignment in the textbook for the next class. The quiz was graded as soon as it was completed and students received immediate feedback consisting of the correct answers and an explanation of the correct answer that included a quotation from the textbook providing the answer along with its page and paragraph location in the textbook. Students were aware from the syllabus and instructions in class that if their online-quiz-score was greater than their exam score, their exam score would be increased.

During the following lecture, at the appropriate moments, each of the In-Class (C) questions was presented as clicker questions. The question was presented as a Power Point slide and the students responded by using personal response devices (clickers). The correct answer was presented immediately after the responses were made. Students were

aware from the syllabus and instructions in class that if their in-class-quiz-score was greater than their exam score, their exam score would be increased.

The eighteen question sets covered material presented on weeks 10 through 14 of the course and the 18 E-questions comprised a unit exam presented during week 14. The same 18 E-questions again appeared on the final exam. The unit exam consisted of the 24 questions that included the 18 (E) questions that were part of the experimental design. During the exam, each question was presented on a Power Point slide, one at a time, for a period of time from 45 to70 seconds based on the number of words in the question, and students responded with clickers. The first half of the exam was the delayed feedback condition. For the first half of the exam, after time ran out, the next question was presented without feedback as to the correct answer. The second half of each exam was the immediate feedback condition. In the second half of the exam, after time ran out, a large green check mark appeared next to the correct answer before the next question was presented. The students had 10 seconds per question to absorb the feedback. After the last question, each question from the first half of the exam was again presented briefly (30 seconds) with the correct answer marked. The only reason that students were given slightly more time while presenting delayed feedback was because they would have to go over the entire question once again whereas in the immediate feedback condition, having just answered the question they only needed time to focus on the feedback being given.

The students knew that questions very similar or identical to the questions on the unit exam would appear on the final exam. So they were highly motivated to attend to the feedback. Furthermore, before the immediate condition the students were told to pay

close attention to each correct answer when it was presented and to imagine that it was the answer they gave. They were told in order to increase their chance of responding correctly to the same question on the final exam they should put all of their effort into learning the correct response rather than trying to remember whatever answer they gave. The same instructions were given before presenting the feedback for the questions in the delayed feedback condition.

Hence, 9 of the 18 E-questions on each unit exam were presented with immediate feedback and the other 9 questions were presented with delayed feedback. The questions presented with immediate feedback to one section of the course were presented with delayed feedback to the other section of the course. Within each condition, the questions were presented in the order in which their topics had been covered during the class.

The final exam was presented to one section three days after the unit exam and to the other section one week after the unit exam. On the final exam, first 18 short-answer questions were presented without feedback. Then, the corresponding 18 multiple-choice questions from the unit exam were again presented. At the end of each question, the students were asked to rate their confidence in the answer they just gave. The students were to rate their confidence level as high, medium or low. Students were told that their confidence rating after each question would influence their score for that question. If they picked low-confidence and got the question right, they gained 1 point. If they got the question wrong they lost 0 points. If they picked medium confidence and got the question right, they gained 2 points. If they got the question wrong they lost 1 point. If they picked high confidence and got the question right, they gained 3 points. If they got

the question wrong they lost 2 points. Hence, the students were highly motivated to select the most accurate confidence ratings.

Students took the final exam in the course lecture hall, proctored by the course instructor and teaching assistant. The final was presented on the Sakai course platform. Students answered the questions on laptops and smart phones. The 18 short-answer and 18 corresponding multiple-choice questions, each followed by a confidence rating, were embedded in a 48 question exam in which 24 short-answer questions, each followed by a confidence rating, were followed by 24 corresponding multiple-choice questions, each followed by a confidence rating. The 6 question pairs that were not part of the experimental design were related to the 6 questions on the unit exam that were not part of the experimental design. Students had 48 minutes to complete the exam. The 48 questions were presented to all students in the same order and students could spend as much time on a question as they wished. However, for the multiple-choice questions the 5 alternatives were presented in a different randomly selected order to each student. So, for each multiple-choice question, each letter, A, B, C, D and E, indicated the correct alternative for one-fifth of the students answering that question. Also, once a question had been answered a student could not go back and change the answer.

4.2 Results

An analysis was performed on the corresponding homework and clicker questions presented before the questions on unit exam that appeared in each condition. On the homework, percent correct was 74% (C.I. = 68, 78) when the questions preceded exam

questions in the immediate feedback condition and 74% (C.I. = 68, 79) when the following exam questions were in the delayed feedback condition. On the clicker questions, percent correct was 75% (C.I. = 68, 83) when the questions preceded the immediate feedback condition and 74% (C.I. = 66, 81) when they preceded the delayed feedback condition. A 2 x 2 analysis of variance in which items was the random factor was performed on the effects of Question type (Homework (H) versus Clicker (C)) and Feedback subset (following exam questions were in delayed feedback condition versus following exam questions were in the immediate feedback condition. The effects of Feedback, $F(1,17) = .6$, $p = .5$, Question type, $F(1,17) = .04$, $p = .85$ and the interactions between Feedback and Question type, $F(1,17) = .5$, $p = .48$ were not significant.

Next, performance on the unit exam was analyzed. Percent correct in the delayed feedback condition was 85% (C.I. = 84, 87) and percent correct in the immediate feedback condition was 89% (C.I. = 87, 90), which was significant in an analysis of variance in which subjects was the random effect, $F(1,384) = 17$. However, when items was the random effect the difference was not significant, $F(1,17) = 1.3$, $p = .277$.

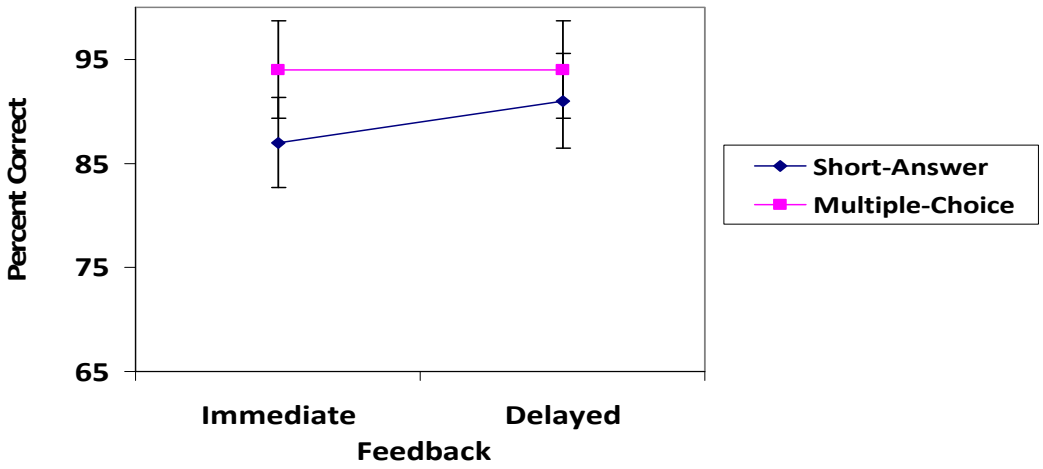

Figure 3: The effect of immediate and delayed feedback on multiple-choice questions on percent correct for subsequent short-answer and multiple-choice questions.

Second, the results of the follow-up exam were analyzed. Since there was not an effect of the retention interval on the results of Experiment 2 this factor was not included in the analysis reported here. However, section was confounded with the specific items answered in each condition by the students in that section and the sections had different numbers of students. This essentially meant that the retention interval between the unit exam and final was confounded with the variable of interest, feedback. Including section resulted in an extremely complex analysis that ultimately produced the results reported here. The results are shown in Figure 3. A 2 x 2 x 2 analysis of variance in which subjects was the random factor was performed on the effects of Feedback on the initial exam (delayed versus immediate) and the Question type (recall versus multiple choice),

which were both within-subjects factors and (course) Section, which was a between-subjects factors. The effects of Feedback, $F(1,383)= 11$, and Question type, $F(1, 383)=140$, and Section were significant, $F(1, 383)=9.7$. Also, the interactions between Feedback and Question type, $F(1, 383)= 27$, and Feedback, Question, and Section, $F(1, 383)= 62$ were significant. As shown in Figure 3, the Feedback by Question interaction is that for short-answer questions percent correct was greater following delayed feedback than following immediate feedback, but for multiple-choice questions there was no difference. The Feedback by Question by Section interaction was that percent correct was greater for the short-answer questions answered by one section following delayed feedback and the other section following immediate feedback (there was no difference for multiple-choicequestions).

Also, a 2 x 2 analysis of variance in which items was the random factor was performed on the effects of Feedback on the initial exam (delayed versus immediate) and the Question type (recall versus multiple choice), which were both within-items factors. The effects of Feedback, $F(1,17) = 2.53$, $p = .13$ and Question type, $F(1,17)=2.89$, $p = .11$ were not significant. However, the interaction between Feedback and Question type, $F(1,17) = 5.43$, was significant. Consequently, the min-F's for Feedback, $F'(1,26) = 2.03$ and Question, $F'(1,18) =2.84$ were not significant. However, the min-F for the interaction between Feedback and Question type, $F'(1,24) = 4.52$, was significant.

Percent correct for high, medium, and low-confidence responses was 96% (C.I. = 96, 97), 86% (C.I. = 84, 91), and 70% (C.I. = 65, 74), respectively for the short-answer questions and 98% (C.I. = 98, 99), 86% (C.I. = 87, 94), and 69% (C.I. = 64, 75), for the

multiple-choice questions. A within-subject analysis of variance in which subjects was the random effect was performed on percent correct categorized by confidence level for the 168 students who used all three confidence levels for both short-answer and multiple-choice questions. The factors were Confidence (high versus medium versus low) and Question (short-answer versus multiple-choice). The effect of Confidence, $F(2,334)= 144$, was significant but the effect of Question, $F(1,167)= 2.29$, was not significant.

Mean confidence for short-answer questions following delayed feedback, for short-questions after immediate feedback, for multiple-choice questions following delayed feedback, and multiple-choice questions following immediate feedback was 2.36 (C.I. = 2.32, 2.41), 2.42 (C.I. = 2.38, 2.46), 2.67 (C.I. = 2.64, 2.70), and 2.68 (C.I. = 2.64, 2.71), respectively. A 2 x 2 within-subjects analysis of variance in which subjects was the random factor was performed on these corresponding confidence ratings on the effects of Question (short-answer versus multiple-choice) and Feedback (delayed versus immediate). The effects of Question, $F(1,384)= 464$ and Feedback, $F(1,384)= 4.8$ as well as their interaction, $F(1,384)= 8.1$ were significant. Also, a 2 x 2 within-items analysis of variance on the effects of Question and Feedback was performed in which items was the random factor. The effect of Question, $F(1,17) = 14$ was significant but the effect Feedback, $F(1,17)= .132$, $p = .721$ and the interaction, $F(1,17) = .306$, $p = .587$ were not significant. Consequently, when min-F's were computed, the effect of Question, $F'(1,19) = 13.6$ was significant but the effect Feedback, $F'(1,19) = .13$ and the interaction, $F'(1,19) = .29$ were not significant.

The correlation between confidence ratings and percent correct was $r(17) = .90$ for multiple-choice questions following delayed feedback and $r(17) = .92$ for multiple-choice questions following immediate feedback. The correlation between confidence ratings and percent correct was $r(17) = .61$ for short-answer questions following delayed feedback and $r(17) = .76$ for short-answer questions following immediate feedback.

Median probability of answering the following corresponding multiple-choice question when the short-answer question was answered correctly was 99%.

4.3 Discussion

As was the case for Experiments 1 and 2, percent correct did not differ between the homework and classroom questions that preceded the corresponding exam questions that received delayed feedback and the corresponding exam questions that received immediate feedback. So before the exam there was no evidence that the students had different levels of knowledge of the fact statements answering the questions for which they received delayed feedback versus those for they received immediate feedback on the exam.

As was the case with Experiment 1, performance on the unit exam was slightly better for immediate feedback than for delayed effect. This effect was significant for subjects but not for items.

As mentioned in the results section, we were forced to confound the retention interval between the unit and final exam with feedback, which was the variable of interest. Analyzing the results with section as a between-subjects factor showed that for

the section with the longer retention interval (1 week), students did significantly better on the short-answer questions for which delayed feedback had been provided on the initial exam than for questions for which immediate feedback had been provided, but students in the section with the shorter retention interval (3 days) did not do better on recall questions that were in the immediate feedback condition on the unit exam. There was no effect on multiple-choice questions. These results were still consistent with those of Experiment 2 where the shortest retention interval was one week. Furthermore, Butler, et al. (2007) found that delayed feedback led to better final test performance relative to immediate feedback over the longer retention interval of one week but did not find significant effects over the shorter retention interval of 1 day. These results indicate the possibility that the benefit of delayed feedback may require longer periods of time to emerge. Hence, retention interval may have a relationship with the effect of delayed feedback on recall questions and this issue certainly warrants further investigation.

Overall, the results of Experiment 2 were replicated, and students did better on the short-answer questions for which delayed feedback had been provided on the initial exam than for questions for which immediate feedback had been provided. Students did not do better on the multiple-choice questions for which delayed feedback had been provided on the initial exam.

Furthermore, a detailed examination of the responses to questions that students got wrong on both the unit and final exams confirmed that there was no difference in performance on the multiple-choice questions as the result of the type of feedback. As shown on the third row of Table 2, on the unit exam, a student averaged 1.3 errors in the

delayed feedback condition and .98 errors in the immediate feedback condition. On the final exam, for those questions that had been in the delayed feedback condition that the students again got wrong, they selected their previous wrong answer 10% of the time. For those questions that had been in the immediate feedback condition that the students again got wrong, they selected their previous wrong answer 8% of the time. Since these were five-alternative multiple-choice questions, the implication of selecting the same wrong response less than 20% of the time when they again got the question wrong is that the students did not remember the correct answer from the feedback on the unit exam and sometimes did not remember, hence avoid, their previous wrong answer.

Confidence ratings were accurate. Higher confidence ratings were associated with a higher percentage of correct responses. Confidence ratings were higher for multiple-choice questions than for short-answer questions. Delayed versus immediate feedback on the unit exam had no effect on confidence ratings for multiple-choice questions on the final exam. Immediate feedback on the unit exam resulted in slightly higher confidence ratings for short-answer questions on the final exam. This effect was significant over subjects but not over items. Furthermore, even though immediate feedback slightly increased confidence on subsequent short-answer questions, delayed feedback significantly increased percent correct for the short-answer questions. Since confidence was **lower** following delayed feedback, the improved performance following delayed feedback cannot be an effect of an **increase** in confidence.

Hence, the effect of delaying the feedback is to increase the probability of generating the correct answer but not the probability of recognizing it. Presumably,

delaying the feedback has this effect because the delayed feedback acts as a cue that causes the student to recall, i.e., to generate the initial question and answer. The distributed generation of the answer increases the probability that it will again be generated after a retention interval.

5 General Discussion

First the effect of immediate feedback on exam performance will be discussed. Then the effect of delayed feedback on subsequent exam performance will be discussed.

In two of three experiments, percent correct on an exam was slightly better when feedback was given after each question. This effect was significant over subjects but not items. So, immediate feedback improves performance for only a few of the items. Butler et al. (2008) found that feedback increases confidence for correct, low-confidence responses on a subsequent exam. Perhaps the effect of immediate feedback on subsequent questions on the same exam is the same as for questions on a subsequent exam: it is to increase the probability of selecting a correct, low-confidence, response to a multiple-choice question. If this was the case then it only affected some questions, as indicated by failure to achieve significance when items was the random effect. It remains to be determined what question characteristics mediate the effect of immediate feedback or lack thereof.

In two experiments, delayed feedback on multiple-choice questions improved performance on subsequent short-answer versions of those questions. In contrast, in three experiments, delayed feedback had no effect on performance on a follow-up presentation of the multiple-choice questions. The interaction between question type and feedback was significant over both subjects and items. This extremely conservative analysis provides strong evidence of its reliability. Furthermore, Experiment 2 found the interaction over two different retention intervals and Experiment 3 demonstrated that it was not mediated by an effect on response confidence. The interaction is consistent with

results collected over the last 60 years and organizes what had previously appeared to be contradictory findings because the different effects on short-answer versus multiple-choice questions had not been noticed.

The specific effect of delayed feedback on short-answer questions is an important finding of theoretical significance because it is specifically predicted by the dual-system model of memory and the generate-and-recognize model of recall that may be derived from it. Within the context of the dual-system hypothesis delayed feedback retrieves the correct answer through the habit system. Hence, the subsequent probability of the habit system generating the correct answer is increased. Consequently, probability of generating the answer as a response to a short-answer question is increased. On the other hand, a multiple-choice question does not require the student to generate an answer so delayed feedback does not affect subsequent performance on this task.

Furthermore, detailed analysis of the results is completely consistent with the hypothesis that the correct answer is first generated and then recognized for the short-answer questions but only recognized for the multiple-choice questions. If the answer must be recognized in order to be the response to the short-answer question then the same answer should again be recognized when included as a possible response to a subsequent multiple-choice question. In fact, as mentioned above, when the short-answer question was answered correctly, the median probability over items that the subsequent corresponding multiple-choice question was answered correctly was 99%. Also, if multiple-choice questions are recognition questions then percent correct should be highly correlated with confidence in the response since these are two different measures of

recognition. In fact, as mentioned above, the correlation between percent correct and confidence for multiple-choice questions was $r(17) = .90$ following delayed feedback and $r(17) = .92$ following immediate feedback. In contrast, if percent correct for short-answer questions is influenced both by generation and by recognition then the correlation between percent correct and confidence should be reduced. In fact, as mentioned above, the correlation between percent correct and confidence for short-answer questions was .only $r(17) = .61$ following delayed feedback and only $r(17) = .76$ following immediate feedback.

That the experiment was embedded in a course insured that the students would be highly motivated participants in the experiment. Hence, there can be no doubt that they intensely attended to both the immediate and delayed feedback for exam questions they knew would be repeated on the final exam. Similarly, they devoted maximum effort to answering all questions correctly. Therefore, compared with a mere laboratory study on which the student's grade did not depend, the course embedded paradigm seems to have very high levels of subject compliance and performance, hence may be the best possible test of a theoretically or pedagogically motivated prediction.

The positive effect of delayed feedback on subsequent short-answer questions also has an important implication for pedagogy. It demonstrates that one does not have to include short-answer questions in a lesson plan to improve performance on them. Distributed presentation of feedback for multiple-choice questions may more efficiently perform the same function.

Table 2: The average number and percent of errors on the unit exam and the average percentage of times a student selects the same wrong alternative on the final exam.

Experiment	Number of questions on unit exam	Mean number of incorrect responses on unit exam		Mean percent of incorrect responses on unit exam		Mean percent of incorrect responses repeated on final exam		t test
Feedback		Delayed	Immediate	Delayed	Immediate	Delayed	Immediate	
1	68	5.3 (4.9,5.6)	4.7 (4.3,5)	18 (17,19)	16 (14,17)	23 (21,26)	20 (18,22)	t(366)= 2.2, p=.03
2	36	5.2 (4.8,6)	5.17 (4.6,6)	29.2 (25,33)	29 (25,32)	19 (13,25)	15 (10,21)	t(33)=. 88, p=.4
3	18	1.3 (1.2,1.4)	0.98 (.86,1,1)	15 (13,16)	11 (10,12)	10 (7.5,11.5)	8 (4.7,11.7)	t(150)= 1.2, p=.23

Note. Confidence intervals are reported with all the means and the t test gives the difference between mean repeated incorrect responses following delayed versus immediate feedback.

References

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.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. III (2007). The effect of type and timing of feedback on learning from multiple choice tests. *Journal of Experimental Psychology: Applied*, 13, 273-281. doi: 10.1037/1076-898X.13.4.273

Butler, A. C., Karpicke, J. D., & Roediger, H. L. III (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 34, 918-928. doi: 10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L. III (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, 36, 604-616. doi: 10.3758/MC.36.3.604

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335-339. doi:10.1016/S0022-5371(73)80014-3

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology*, 29, 831-848. doi: 10.1080/01443410903310674

Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language*, 52, 595 - 617. doi:10.1016/j.jml.2005.01.015.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, 58, 79–97. doi: 10.3102/00346543058001079

Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neuropsychology of Learning and Memory*, 65, 65-72. doi:10.1006/nlme.1996.0007

Pressey S. L. (1950). Development and appraisals of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology*, 29, 417-447. doi:10.1080/00223980.1950.9916043

Pressey, S. L. (1926). A simple apparatus which gives tests and scores and teaches. *School and Society*, 23, 373–376.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Saunderson, A. (1974). Effect of immediate knowledge of results on learning. *Australian Mathematics Teacher,* 30, 218-221.

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. doi: 10.1111/j.1467-9280.1992.tb00029.x

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, 24, 86-97. doi:10.1037/11324-010

Sullivan, H. J., Schultz, R. E., & Baker, R. L. (1971). Effects of systematic variations in reinforcement contingencies on learner performance. *American Educational Research Journal*, 8, 135-142. doi: 10.3102/00028312008001135

White, K. (1968). Delay of test information feedback and learning in a conventional classroom. *Psychology in the Schools*, 5, 78–81. doi: 10.1002/1520-6807(196801)5:1<78::AID-PITS2310050113>3.0.CO;2-Q

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* 7, 464-476. doi:10.1038/nrn1919

Appendix

Experiment 1:

H. Students studied student ID pictures and were tested on both the same pictures and drivers license pictures of the same students. Half of the pictures were of classmates and half of the pictures were of strangers. Which test condition or conditions produced the poorest recognition?

A. ID and driver's license pictures of the unfamiliar students.

B. Recognition performance was perfect or almost perfect in all conditions.

C. Driver's license pictures of unfamiliar students.

D. Driver's license pictures of both classmates and unfamiliar students.

E. Recognition performance was less than 90% in all conditions.

C. Students are shown a set of famous faces of the 21st century and famous faces of the first decade of the 20th century. The test consists of either the same or a different picture of the same face as a target. The students were poorest at recognizing

A. New pictures of 20th century famous faces.

B. New pictures of both 20th and 21st century famous faces.

C. Old and new pictures of the 20th century famous faces.

D. Recognition performance was perfect or almost perfect in all conditions.

E. Recognition performance was less than 90% in all conditions.

E. Students are shown a study list consisting half of pictures from their high school yearbook and half of pictures from another high school yearbook. Test items are pairs of pictures, one of which was a person whose picture was shown on the study list. Half the target pictures are repeated yearbook pictures and the other half of the target pictures are driver's license pictures of the same people. Which test condition or conditions produced the poorest recognition?

A. Yearbook and driver's license pictures of the unfamiliar high school students.

B. Driver's license pictures of both classmates and unfamiliar students.

C. Driver's license pictures of unfamiliar students.

D. Recognition performance was perfect or almost perfect in all conditions.

E. Recognition performance was less than 90% in all conditions.

All three questions above were verified by the following statement in the reference page:

"The observers were virtually perfect at recognizing faces of familiar individuals whether the same picture or a different picture was shown. The observers were also virtually perfect at recognizing exactly the same picture of the face of a stranger. However, recognition of the face of a stranger from a different picture was only slightly above chance."

Experiment 2:

H. Which spelling error is least likely to be detected?

A. Siller

B. Firn

C. Selter

D. Saller

E. Farn

C. Which spelling error is most likely to be detected?

A. werk

B. wark

C. Soller

D. wurk

E. wirk

E. Which spelling error is most likely to be detected?

A. Furn

B. Ferm

C. Furm

D. Forn

E. Nosion

E_s. Because it is not a homonym of a word, the non-word forn is _____ to be detected

All four questions above were verified by the following statement in the reference page:

“Homophones are two different letter sequences that are pronounced the same way, e.g., cellar and seller, work and werk. When only one homophone is a word the visual whole-word and auditory letter-sequence pathways produce conflicting responses to the input. In a spelling-error detection task a homophone substitution such as werk for work was less likely to be noticed than a nonhomophone substitution such as wark for work (Corcoran, 1966; 1967; Corcoran & Weening, 1968; Mackay, 1968). “

Experiment 3:

H. Retrograde amnesia may result from

A. A blow to the head

B. Hypnosis

C. Smoking

D. Emotional shock

E. Obesity

C. Which is an effective treatment for retrograde amnesia that results from a blow to the head?

A. Another blow to the head

B. Hypnosis

C. Alcohol

D. Electric shocks

E. None of the above

E. Retrograde amnesia from a blow to the head may be treated by

A. hypnosis

B. psychoanalysis

C. Valium

D. another blow to the head

E. none of the above

E_s. Retrograde amnesia from a blow to the head _____ be treated by another blow to the head

All four questions above were verified by the following statement in the reference page:

“Usually, retrograde amnesia is the result of bilateral injury to the medial temporal cortex. If a person receives some kind of shock to the brain, such as a severe blow, he or she may forget events that occurred during some time period leading up to the moment of the trauma. Though no longer oriented to place and/or time, the person remembers his or her identity. In general, the more severe the shock, the longer is the time period that is forgotten. Thus football players who are stunned by a hard tackle may forget a few seconds of their lives. But a patient who receives electroconvulsive shock treatment (ECT) in a mental hospital or a survivor of a severe auto accident with a major skull injury may forget months or even years.

When the memories of a person suffering from severe retrograde amnesia begin to return, the pattern in which they do so is quite disorganized. At first only a few memories are recovered, and the person may be unable to place them in the right temporal order. Two separate events may be combined into one. As more and more events are recalled, the person is able to create islands of remembering; that is, a series of related events may be placed together in their correct chronological order. As more events are recalled, the islands become bigger and the gaps between them become smaller, until finally the islands merge and the complete episodic record is restored. In rare cases a large temporal gap never closes, and the person in effect loses a few years from his or her life. ”