The Combined Effects of Feedback and Delaying Initial Retrieval on Learning

Nicola Marie Crane
Lancaster University

Running head: THE IMPACT OF FEEDBACK AND DELAY ON LEARNING OF PROSE MATERIALS

The Combined Effects of Feedback and Delaying Initial Retrieval on Learning

Nicola Marie Crane

2009

A dissertation submitted to Lancaster University in partial fulfilment of the requirements for the degree of BSc (Hons) in Psychology

The work submitted in this report is my own and has not been submitted in substantially the same form towards the award of another degree or other qualificatory work by myself or any other person. I confirm that acknowledgement has been made to assistance given and that all major sources have been appropriately referenced.

Nicola Marie Crane

6th February 2009

Abstract

Giving individuals a test on information which s to be learned has been found to be beneficial for long-term retention, as has enforcing a delay between exposure to the information and the initial tets. However, the benefits of delaying initial retrieval have not been found when the information comes from prose materials. The present experiment aimed to demonstrate that delaying initial retrieval can be beneficial for the learning of prose materials if feedback is given on test answers. Forty participants read a passage of text and took a multiple-choice test after either a two or ten minute delay. Half of participants received feedback and half did not. Participants returned 24 hours later to complete a final test to assess their retention of information. Feedback was found to have a beneficial effect on retention, however delay had no effect. It was concluded that a ten minute delay was not long enough to enhance learning when using multiple choice testing and that combining multiple-choice testing with immediate feedback can be detrimental for learning.

The Combined Effects of Feedback and Delaying Initial Retrieval on Learning

In the exploration of memory improvement, many claims have been made that certain factors are useful for enhancing long-term retention. These are not exclusively the product of modern research; many techniques were created in ancient times, for example, mention of mnemonic link systems such as the method of loci can be found in texts by Aristotle (Sorabji, 1972). Recent studies, along with older ideas, have produced what are now well-known techniques used by both students and teachers to aid with the learning and retention of information.

The use of many of these techniques often provides a noticeable increase in both the amount of information recalled and the amount of information that the learner believes that they will be able to recall at a later time. It has been found that predictions made by an individual on the likelihood of later recall are sometimes accurate Blake (1973), but in many cases do not reflect the actual chance of later remembering (Agarwal, Karpicke, Kang, Roediger, & McDermott, in press; Roediger & Karpicke, 2006a). Indeed, some research has found that counterintuitive methods, the use of which may lead the learner to feel that their recall has been impaired, can actually increase the chance of later recall. One example is the testing effect. After initial exposure, the learner will show superior long-term retention if given a test on what they have learned, as opposed to being allowed to review the material again instead (Carrier & Pashler, 1992; Agarwal et al., in press; Roediger & Karpicke, 2006a). The length of time between the learning session and first test can play an important part in learning too, with a longer delay reducing recall on a short-term basis, but enhancing long-term retention (Modigliani, 1976; Karpicke & Roediger, 2007). Although the effectiveness of delaying initial retrieval has been demonstrated using stimuli such as word pairs, it has not been found when prose materials have been used, argued by Karpicke and Roediger (2007) to be a consequence of the difficulty of the material itself. The present experiment was carried out in order to investigate whether delaying initial retrieval can benefit the retention of prose materials when feedback is used to counteract any forgetting.

The testing effect is not a newly-discovered phenomenon. In a study involving 3605

participants, Spitzer (1939) demonstrated the strength of the testing effect. Individuals were tested once, twice or three times after learning, between 0 and 63 days after the original learning session. Those who were given a test after learning performed much better on a final test than those who were not given an intervening test, no matter whether the initial test was 1 day or 21 days after learning. Although this research highlighted the importance of the practical uses of testing for enhancing learning as well as assessing it, little consideration was given to why this effect actually occurs.

Recent research has attempted to provide a theoretical explanation for the testing effect (for a review, see Roediger & Karpicke, 2006b). Bjork (1994) argues that the testing effect works as it provides "desirable difficulty" for the learner that makes initial retrieval difficult, but enhances later retrieval. Schmidt and Bjork (1992) argue that creating desirable difficulties encourages the learner to use varied processing strategies, increasing the likelihood of transfer-appropriate processing (Morris, Bransford, & Franks, 1977). When the learner attempts to retrieve the information, this retrieval will use more complex processes in memory than if the learner is simply re-exposed to the material. If the retrieval strategies are consistent with those used on an earlier attempt, there is a greater chance of success, as the learner should find it easier to use these retrieval strategies again in the future.

Research by Gardiner, Craik, and Bleasdale (1973) supports Bjork's (1994) theory of desirable difficulties, but gives a different explanation of why they are effective in enhancing long-term retention. In this research, participants were given the first letter and definition of an uncommon word and asked to guess what the word was. Afterwards, participants were asked to write down as many of these words as they could remember. It was found that the longer participants took to think of a word on the initial test (with time presumed by Gardiner et al. to reflect retrieval difficulty), the more likely it was that the word would be recalled later. Gardiner et al. argue that this is because of different types of retrieval during the initial generation of the word and its subsequent recall. The initial retrieval of the word is from semantic memory, but the later retrieval is from episodic memory. When a word which takes longer to recall is processed, more peripheral information (such as the meaning of the word, and similar words) is activated. This detailed retrieval attempt results in greater chance of later retrieval from

episodic memory.

Further evidence supporting Bjork's (1994) 'desirable difficulties' theory has been drawn from research exploring expanding retrieval practice. Expanding retrieval practice (Landauer & Bjork, 1978) is a technique which involves multiple retrieval attempts with increasingly larger gaps between them, believed to a highly effective method for increasing the likelihood of later retention. Until recently, there was little evidence to contradict this theory. However, Karpicke and Roediger (2007) reviewed numerous experiments investigating expanding retrieval practice and found that they frequently confounded the spacing of the intervals with the delay before the first retrieval attempt. In the same paper, Karpicke & Roediger reported the results of experiments in which they had compared the effects of expanding and equally spaced retrieval without this confound. They found little difference between the two types of spacing, with equally spaced retrieval actually being more effective than expanding retrieval in improving later retention of information. More importantly, however, they concluded that the timing of the first attempt to retrieve information after encoding is crucial to later retention, with a longer delay producing improved long-term retention.

Delayed initial retrieval is believed to work in a similar way to the testing effect in that creating difficulty for the learning is detrimental for short-term retention, but is beneficial on a long-term basis. Balota, Duchek, and Logan (2007) argue that this is because of encoding variability. For information to be remembered after learning, there must be similarities between contextual elements in the original learning and later retrieval attempts. It is more likely that these contextual elements will have changed to some degree when there has been a delay (compared to when there has been no delay, or a very short delay) and so a wider variety of contextual elements will increase the likelihood of later recall.

Roediger and Karpicke (2006a) gave further evidence of the benefits of delayed initial retrieval by either asking participants to study a passage of text twice, or to study it once and then take a free recall test on the text. A second free recall test was given after an interval of either five minutes, two days or one week. Roediger and Karpicke found that after a five minute delay, participants who had studied the passage of text twice performed better than those whose initial study session had been followed by a test. However, after a two day or one week interval,

these results were reversed, with participants recalling more information if they had been tested on the material learned, rather than restudying it. Other research has shown similar results (e.g., Carrier & Pashler, 1992; Cull, 2000; Agarwal et al., in press) and demonstrated how testing is an active process which modifies the memory trace more than the fairly passive process of rehearsal. This previous research also highlights how the testing effect is not merely due to the repeated exposure to the information, which is experienced during retrieval in an initial test, and shows how testing produces better long-term retention than does rehearsal of the material to be learned. Although there is no conclusive evidence on which account best explains the theoretical underpinnings of the effects of testing or delayed initial retrieval, Roediger and Karpicke (2006b) point out that these theories may be viewed as complementary to each other, and not mutually exclusive. However, there is a general consensus that making the initial retrieval more difficult for the learner results in superior long-term retention.

There are limitations to the advantages of delaying initial retrieval, and making retrieval more difficult is not always beneficial. Karpicke and Roediger (2007) note that the positive effects of delaying retrieval can be lost when the delay is so long that the learner is unable to retrieve the relevant information. There is some evidence supporting this claim, from research in which the stimuli used consisted of prose passages. When testing participants on recall of information from prose materials, Spitzer (1939) found that an immediate initial test produced better subsequent recall than if the test was delayed by 1 day or more. Karpicke & Roediger (2006, as cited in Karpicke & Roediger, 2007) found similar results when comparing immediate testing to tests delayed by just 10 minutes. Karpicke and Roediger (2007) speculate that this is because of the complex nature of prose and that the amount of forgetting is sufficient to negate any benefit of delayed retrieval. However, they argue, if participants are given feedback, a longer delay before the initial retrieval attempt may result in better long-term retention. The testing effect, in research using word pairs, has been found to be stronger when feedback is given (Cull, 2000; Agarwal et al., in press) and the same trend has been found in research focusing on delayed initial retrieval (Karpicke & Roediger, 2007). Previous research concerning the combination of feedback with delayed initial retrieval has mainly used stimuli such as word pairs (e.g. Karpicke & Roediger, 2007). The present experiment aims to extend the previous

body of work by demonstrating the delayed initial retrieval effect with prose materials, using feedback to counteract forgetting.

Various types of stimuli have been used when investigating delayed initial retrieval and the testing effect. Recent research documents the testing effect using different stimuli such as word-pairs (Jacoby, 1978), nonsense syllables (Carrier & Pashler, 1992), prose passages (Agarwal et al., in press) and facts taken from educational materials (Carpenter, Pashler, & Cepeda, in press), all with similar results. In terms of test format, multiple-choice tests have been found to be particularly useful as participants responses are either correct or incorrect and there is no need for more subjective interpretations of responses. Also, unlike with short answer tests, responses can be checked automatically by a computer and so feedback can be given to participants as desired by the experimenter, without needing to take the time to assess participants answers. The simplicity of both constructing and assessing multiple-choice question tests can also be seen as advantageous when considering the practical applications of testing, in an educational setting.

Multiple-choice tests are not without their problems though; Butler and Roediger (2008) note that as a result of being exposed to realistic but incorrect answers to choose from, learners may come to believe that some or all of this false information is true. However, the use of feedback can be useful: Butler and Roediger (2008) gave participants two multiple-choice tests and found that giving feedback significantly reduced the chance of participants choosing an incorrect answer on the second test after also choosing it on the first test. Bangert-Drowns, Kulik, Kulik, and Morgan (1991) conducted a review of methodology in research which involved the use of testing and feedback. They calculated the average effect size of feedback in the different studies and found that individuals who completed multiple choice tests benefitted the most from feedback, compared with completion tests, short answer tests and tests which involved a combination of these. They also found a much greater effect of feedback when the correct answer was provided in comparison with when participants were simply told that they were correct or incorrect. Feedback aids learning because it facilitates error correction and allows learners to identify any incorrect knowledge they have or gaps in their knowledge (Bangert-Drowns et al., 1991). Butler and Roediger (2008) argue that it also gives the learner confirmation about correct answers and so helps to maintain this knowledge.

The current research used multiple-choice tests for the theoretical reasons discussed above, as well as for practical reasons. The aim of the current research was to examine the effect of combining feedback with a delayed initial retrieval attempt. After reading a passage of prose text, participants completed a distractor task for either two or ten minutes. Participants then completed a multiple-choice question test, either with or without feedback. If feedback was displayed, as well as being told if their response was correct or incorrect, participants were shown the correct answer. This was done as it has been found to increase the effectiveness of the feedback (Bangert-Drowns et al., 1991). Participants returned 24 hours later to complete the final multiple-choice test. Participants scores on each test were recorded, along with response times, which can be used to assess the amount of difficulty experienced by participants (Gardiner et al., 1973). A ten minute delay was thought to be a sufficiently length to encourage effortful retrieval as previous research by Karpicke & Roediger (2006, as cited in Karpicke & Roediger, 2007) found that participants tested on prose materials after this length of time recalled less than participants tested after a much shorter delay, indicating greater difficulty in retrieval.

On the initial test, it was expected that there would be no effect of feedback, but there would be an effect of delay. Participants who took the initial test after a ten minute delay were expected to show worse performance than those who took this test after a two minute delay, due to forgetting.

Regarding the final test, it was hypothesized that there would be an interaction between feedback and delay. A ten minute delay should result in more effortful retrieval on the initial test than a two minute delay. Giving participants feedback on their responses should lead to any errors due to forgetting being corrected and so, participants initially tested after a ten minute delay and given feedback were expected to have higher scores on the final test than participants tested after a short delay who were given feedback. Participants who were not given feedback were expected to show superior performance on the final test when the initial test was after a two minute delay, rather than a ten minute delay, as was found by Karpicke and Roediger (2007), due to the prose material being complex enough for forgetting to overcome any benefits of delayed initial retrieval.

It was deemed theoretically relevant to see if there was any correlation between response

times on the initial test and score on the final test. For all participants who received feedback, it was expected that there would be a positive correlation between these factors, with longer response times leading to higher scores on the final test. However, for participants who did not receive feedback, it was expected that there would be no correlation between initial retrieval difficulty and subsequent test scores. This is because delaying initial retrieval should lead to "desirable difficulties" in recall (Bjork, 1994) which should enhance later retrieval. However, due to the complexity of prose material, forgetting which is not counteracted by feedback should counteract the beneficial effects of delayed retrieval (Karpicke & Roediger, 2007).

Method

Participants

The participants were 40 undergraduate students from Lancaster University, all aged between 18 and 22 years, selected using opportunity sampling. All participants' first language was English. Prior to the start of the experiment, participants were told what it involved, so that they could give informed consent. Participants were informed that they could withdraw from the experiment at any time and for any reason. After the experiment, participants were fully debriefed on the aims and rationale behind the experiment.

Design

A 2x2x2 between-participants factorial design was used. Two of the independent variables, feedback and delay length were manipulated between-participants. Feedback had two levels (feedback given during the first test, or no feedback at all) and delay had two levels (two minutes or ten minutes). The other independent variable, test session (i.e. test in the first session and test in the final session) was a within-participants factor.

The dependent variables were the number of questions which were answered correctly in each test, and the amount of time it took participants to answer the questions in each test.

Apparatus and Materials

The short passage of text was 586 words long and was taken from Sports Heroes, Fallen Idols by Stanley Teitelbaum (see Appendix A). The passage of text described events in the 1940s in America involving two American football players who were accused of accepting bribes. It was chosen as it was a topic that participants were unlikely to be familiar with.

In the delay phase, participants played a problem-solving puzzle game called Peggle. This was thought to be an adequate distractor task as the game requires concentration and so participants should not be able to recall information about the passage of text from working memory.

Computer software called jPsyscript was used to display the questions. This is a computer programming language designed to aid in the construction of psychology experiments. It is based on PsyScript computer software developed at Lancaster University. The multiple-choice question test was designed using key facts from the passage of text. The decoy answers needed to be realistic choices and so were generated using information from an internet encyclopaedia. There were twenty questions in total, each with five possible answers; one correct answer and four decoy answers.

Procedure

In the first session, participants were told that the experiment would involve reading a passage of text, playing a computer game and then completing a multiple choice question test. Participants were also asked if they had read the book from which the passage of text was taken, and if they had any interest in American football. None of the participants answered yes to either of these questions. They were told that they had the right to withdraw at any time and that their results would remain anonymous. Participants were randomly assigned to the long or short delay condition and feedback or no feedback condition. They were then given the passage of text to read. When they had finished reading it, they played a computer game, Peggle, for either two minutes or ten minutes, depending on which condition they had been assigned to. When the time was up, they completed the first multiple choice question test of twenty questions. Each question was displayed on screen and the correct answer and

four decoys were shown. When participants clicked on their chosen answer, the next screen was shown. In the 'feedback condition, when participants clicked on their chosen answer, either the word 'Correct' or 'Incorrect' appeared followed by 'The answer was:' and the correct answer was displayed. This screen appeared for three seconds before proceeding to the next question. In the no feedback condition, a blank screen was displayed for three seconds, and then the next question appeared. The pause in the no feedback condition was there to make the multiple-choice question test the same length in each condition. The time between the question being displayed and the participant clicking on an answer was recorded for each question. All questions and answers were displayed in a random order to prevent participants remembering an answer based on its position on the page or order in the sequence of questions. When the test had finished, participants were thanked for their participation.

In the second session, which was 24 hours after the first test, participants were told that they would be given a multiple choice question test. They were informed that the test would be the same as the one completed the previous day, but participants in the 'feedback' condition were informed that they would not receive feedback on their answers and there would be a three second pause with a blank screen between each question. All participants were told that the questions and possible responses would be in a random order and not the same as they were in the first session. Again, response time was recorded. Once the experiment was complete, participants were thanked for their participation and debriefed.

Results

Scoring

Test scores were calculated by totalling the total number of correct responses for each participant. The maximum possible score was 20. Response times were calculated using the average time it took each participants to answer a question, from it being displayed on the screen to the participants selecting their response.

Any result reported significant below is significant at the alpha level of $p < 0.05$.

Overall results

Table 1 shows the scores on both the initial and final tests. A 2x2x2 mixed factorial ANOVA with feedback and delay as between-participants factors and test as a within factor was conducted to analyse the effect that feedback and delay had on scores over both tests. There was no three-way interaction between test, feedback and delay, $F(1, 36) = 0.12$, $p = 0.73$, $\eta_p^2 = 0.01$.

Table 1: Mean number of correct responses on each test, with standard deviation in parentheses

	Test	
Condition	Test 1	Test 2
Feedback given		
Two minute delay	9.80 (3.43)	13.60 (2.84)
Ten minute delay	11.10 (1.85)	15.80 (2.20)
No feedback given		
Two minute delay	12.40 (2.41)	11.70 (2.41)
Ten minute delay	10.60 (2.72)	11.20 (2.82)

An additional 2x2x2 mixed factorial ANOVA was conducted to analyse response times which can be found in Table 2. Again, no three-way interaction was found, $F(1, 36) = 0.09$, $p = 0.77$, $\eta_p^2 = 0.01$.

Lower-level interactions and main effect analyses for both test scores and response times on each test are discussed below to avoid repetition.

Initial Test

A 2x2 between-participants ANOVA was conducted to analyse the effects of feedback and delay on the initial test scores. At this stage, there was no theoretical reason that feedback should have provided any advantage, and this was reflected in the results which found no main effect of feedback, $F(1, 36) = 1.56$, $p = 0.22$, $\eta_p^2 = 0.04$. Unexpectedly, however, no main effect of delay was found, $F(1, 36) = 0.09$, $p = 0.77$, $\eta_p^2 = 0.01$. There was no significant interaction between feedback and delay, $F(1, 36) = 3.38$, $p = 0.07$, $\eta_p^2 = 0.08$.

Participants' response times on the initial test were also analysed. A 2x2 between-participants ANOVA revealed that there was no significant main effect of delay ($F(1, 36) = 1.25$, $p = 0.27$, $\eta_p^2 = 0.03$). There was, however, a significant main effect of feedback ($F(1, 36) = 6.99$, $p < 0.05$, $\eta_p^2 = 0.16$), with participants who had received feedback answering questions faster

Table 2: Mean response time on each test (in milliseconds), with standard deviation in parentheses

	Test	
Condition	Test 1	Test 2
Feedback given		
Two minute delay	6824 (1488)	5819 (473)
Ten minute delay	6440 (744)	5757 (1259)
No feedback given		
Two minute delay	7222 (1715)	5047 (928)
Ten minute delay	8757 (2200)	7160 (1912)

($M = 6632$ms) than participants who were not given feedback ($M = 7990$ms). There was no interaction between feedback and delay, $F(1, 36) = 3.49$, $p = 0.07$, $\eta_p^2 = 0.09$.

Final Test

A 2x2 between-participants ANOVA was used to compare participants' scores on the final test. A significant effect of feedback was found, $F(1, 36) = 15.86$, $p < 0.01$, $\eta_p^2 = 0.31$, with participants who initially received feedback scoring higher on the final test ($M = 14.70$) than participants who did not receive feedback on the first test ($M = 11.45$). There was no significant effect of delay, $F(1, 36) = 1.09$, $p = 0.31$, $\eta_p^2 = 0.03$, and no interaction between feedback and delay, $F(1, 36) = 2.74$, $p = 0.11$, $\eta_p^2 = 0.07$.

Again, a 2x2 between-participants ANOVA was conducted on participants' response times. There was no significant main effect of feedback, $F(1, 36) = 0.63$, $p = 0.43$, $\eta_p^2 = 0.02$. Delay had a significant main effect, $F(1, 36) = 6.65$, $p < 0.05$, $\eta_p^2 = 0.16$, with participants who took the initial test after a two-minute delay ($M = 5433$ms) answering faster than those who took the test after a ten-minute delay ($M = 6458$ms). A significant interaction was found between feedback and delay, $F(1, 36) = 7.48$, $p < 0.01$, $\eta_p^2 = 0.17$. Further analyses were conducted and a one-way ANOVA revealed that there was no significant difference between response times when participants took the initial test after a ten-minute delay, $F(1, 19) = 3.76$, $p = 0.07$, $\eta_p^2 = 0.17$. However, participants whose initial test was after a two-minute delay answered questions faster if not given feedback compared to those who were given feedback, $F(1, 19) = 5.49$, $p < 0.05$, $\eta_p^2 = 0.23$. Additionally, a second one-way ANOVA showed that there was no significant difference in response times of participants who were given feedback, $F(1, 19) = 0.02$, $p = 0.89$, $\eta_p^2 = 0.01$.

However, participants who were not given feedback answered significantly faster on the final test when the initial test was after a two-minute delay than when the initial test was after a ten-minute delay, $F(1, 19) = 9.88$, $p < 0.01$, $\eta_p^2 = 0.35$.

Improvement

A 2x2 between-participants ANOVA was conducted to assess the impact of feedback and delay on the improvement in scores across the two tests, which took into account participants' performance on the initial test, as shown in Figure 1. There was a significant main effect of feedback, $F(1, 36) = 57.28$, $p < 0.01$, $\eta_p^2 = 0.61$, with participants who were given feedback showing a larger increase in score on the final test ($M = 4.25$) than participants who were not given feedback who actually had lower scores on the final test ($M = -0.05$). There was no significant effect of delay, $F(1, 36) = 3.75$), $p = 0.06$, $\eta_p^2 = 0.09$, and no interaction between feedback and delay, $F(1, 36) = 0.12$, $p = 0.73$, $\eta_p^2 = 0.01$.

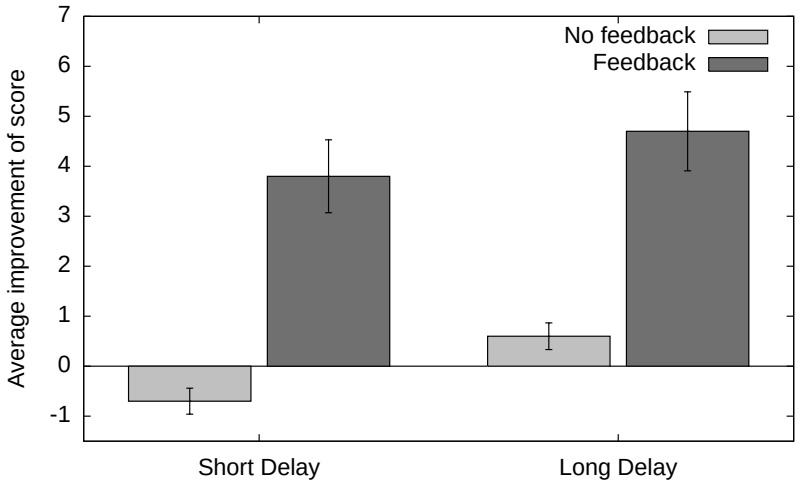

Figure 1. Improvement in scores on the final test. Error bars show standard errors.

The effect of feedback and delay on participants' improvement in response times was also assessed using a 2x2 between-participants ANOVA. There was a significant main effect of feedback, $F(1, 36) = 5.58$, $p < 0.05$, $\eta_p^2 = 0.13$, with participants who were not initially given feedback showing a larger improvement in response time ($M = -1886$ms) than those who were given feedback on the first test ($M = -884$ms). There was no main effect of delay,

$F(1, 36) = 1.04$, $p = 0.31$, $\eta_p^2 = 0.03$. No interaction between feedback and delay was found, $F(1, 36) = 0.09$, $p = 0.77$, $\eta_p^2 = 0.01$.

Initial retrieval difficulty and subsequent performance

A correlation was conducted to test if there was any correlation between response time on the initial test and test score on the final test. Pearson's correlation coefficient was used to test this and for participants who received feedback, no correlation between initial response time and later score was found, $r(18) = -0.05$, $p = 0.83$. No correlation between initial response time and later test score was found for participants who did not receive feedback either, $r(18) = -0.34$, $p = 0.14$.

Discussion

The present research did not support the hypothesis that giving participants feedback would result in better subsequent performance when the initial test was conducted after a ten minute delay than when first tested after a two minute delay. Additionally, the hypothesis that participants who were not given feedback would perform better on a later test when the initial test was conducted after a two minute delay compared to participants first tested after a ten minute delay, was not supported either. Furthermore, no correlation between initial response times and later test scores was found, indicating that there was no relationship between the time it took to answer a question on the initial test and number of questions answered correctly on the final test.

One possible explanation for these results is that the ten minute delay may not have been long enough to encourage sufficiently effortful retrieval. Delay had no significant effect on participants' response times or scores on the initial test, indicating that a similar amount of difficulty was experienced by all participants. Additionally, no correlation was found between intial difficulty and subsequent retrieval. This contradicts the findings of Karpicke & Roediger (2006, as cited in Karpicke & Roediger, 2007) who found that participants who were tested on information from a prose passage showed better long-term retention when the initial test was immediate, rather than after a delay. However, Karpicke and Roediger used free recall tests in

their research, whereas the present research used multiple-choice testing. The crucial difference here is between testing that requires recall and testing that requires recognition. Glover (1989) found that giving participants recall tests resulted in greater improvements in later recall than recognition tests did. Glover argues that this is because recall promotes more complete retrieval than recognition. This seems to indicate that test format, or more specifically, whether the test requires recall or recognition, plays an important role in mediating the effects of delay on subsequent performance.

Further evidence supporting the view that multiple-choice tests may be deficient for testing prose materials comes from Kang, McDermott, and Roediger (2007), who argue that when participants are given feedback on their answers, an initial test which requires recall provides greater benefits for later retrieval than an initial recognition test, even when the final test requires recognition.

It could be argued that an alternative explanation for the failure of the present experiment to support the hypothesis can be seen if the findings are interpreted with encoding variability theory. Balota et al. (2007) argue that enforcing a delay before initial retrieval can allow a change in context and therefore a greater chance in later recall. In the present research, the longer (ten minute) delay before intial testing did not lead to any difference in recall compared with the short (two minute) delay. Therefore, it seems that this delay was not long enough for contextual elements to change sufficiently to aid later retrieval.

An unexpected finding was seen in participants' response times on the initial test which was faster for those who were given feedback compared to those who did not receive feedback. At this stage, feedback had not affected performance, as demonstrated by the lack of significant difference in number of correct answers between the different groups of participants. Therefore, it seems that highly unlikely that these faster response times were caused by retrieval being easier for individuals who had been given feedback.

One possible reason for this result is that participants were aware that they would be returning to complete another test 24 hours later. Participants who received feedback were aware that they would be shown the correct answer regardless of whether or not it matched their response, and also knew that the subsequent test would contain the same material. This idea is

supported by the fact that on the final test, participants who had taken the initial test after a ten minute delay had longer response times (indicative of retrieval difficulty) than those who had taken the test after a two minute delay, regardless of whether or not they had received feedback. This may have approximated the conditions in a study by Anderson, Kulhavy, and Andre (1971) who found that allowing participants to view the answer to a question before actually submitting their own response resulted in poorer learning and subsequently worse long-term retention than participants who received feedback after they had answered each question. Although in the present research superior performance was demonstrated in participants who were given feedback compared to those who were not given feedback, this may have been because of the repeated exposure to correct answers, rather than the correction of their own errors and confirmation of correct responses.

To prevent this problem, future studies should consider the use of delayed feedback. Butler and Karpicke (2007) found that delaying feedback caused a greater increase in later retrieval than giving participants feedback immediately. They argue that this is because immediate feedback creates similar conditions to massed presentation of material, which is detrimental to learning. In light of the findings of the present research, delayed feedback could also be useful for giving the learner more motivation to put effort into remembering the information which is to be learned. Also to be investiagted is the optimal delay between testing and initial retrieval with different test formats. The present results found that a ten minute delay is not long enough for it to be beneficial for learning when an intervening test is in multiple-choice format. Further research should examine if increasing the delay before initial retrieval is beneficial when using multiple-choice tests, and what length delay is optimal, or if the type of retrieval encouraged by these tests is inadequate to fully reap the benefits of delayed initial retrieval.

The present research has an important implication for practical applications of testing in education. Firstly, if individuals are to be given multiple-choice question tests to promote learning, feedback should be used with caution. The present results show that when the learner is aware that they will be tested again on the same materials, giving feedback immediately after each test item led to learners spending less time answering each question. This in turn may have led to shallower learning.

Acknowledgements

Thanks to my project supervisor, Alan Collins, for all the advice and guidance, and for allowing me the freedom to choose pretty much any topic within cognitive psychology that I wanted to. Thanks must also be given to Andrew Brampton for creating jPsyscript for me and for helping me better understand my reasoning by almost constantly questioning why it is that in psychology things are done completely different to how they are done in computer science!

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (in press). Examining the Testing Effect with Open-and Closed-Book Tests. *Applied Cognitive Psychology.*

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology*, *62*, 148-156.

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? a crticial review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L Roediger III.* New York: Psychology Press.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213-238.

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

Blake, M. (1973). Prediction of Recognition When Recall Fails: Exploring the Feeling-of-Knowing Phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *12*(3), 311–319.

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

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

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (in press). Using Tests to Enhance 8th Grade Students Retention of U.S. History Facts. *Applied Cognitive Psychology.*

Carrier, M., & Pashler, H. (1992). The Influence of Retrieval on Retention. *Memory and Cognition*, *20*, 633-643.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory and Cognition*, *1*, 213–216.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392-399.

Jacoby, L. L. (1978). On Interpreting the Effects of Repetition: Solving a Problem Versus Remembering a Solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649-667.

Kang, S. H. K., McDermott, H. K., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 1–31.

Karpicke, J. D., & Roediger, H. L. (2007). Expanding Retrieval Practice Promotes Short-Term Retention, but Equally Spaced Retrieval Enhances Long-Term Retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 704–719.

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

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.

Roediger, H. L., & Karpicke, J. D. (2006a). Test Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*, 249–255.

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

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.

Sorabji, R. (1972). *Aristotle on memory.* London: Duckworth.

Spitzer, H. F. (1939). Studies in Retention. *The Journal of Educational Psychology*, *30*, 641–656.

Appendix A: Prose Material

On December 15, 1946, the night before the championship showdown between the New York Giants and the Chicago Bears, two of the Giants key playersquarterback Frank Filchock and fullback Merle Hapeswere questioned about an allegation that they had been offered a bribe to throw the game. The investigation determined that a small-time gambler named Alvin Paris who socialized with Filchock and Hapes had repeatedly offered each player a payoff of $2,500 and a $1,000 bet on the Bears if they would help throw the championship game. The police had tapped Pariss phone because they suspected his home was a bookmaking establishment. Though they learned that both players had turned down the bribe, neither had reported the attempt to team officials or to the police.

This event was the biggest scandal to hit professional sports since the 1919 Black Sox incident, and the mayor of New York City, William ODwyer, took part in questioning the players. Filchock denied that he had been asked to fix the game, and ODwyer, impressed with his sincerity, believed him. The mayor also thought there was a lack of evidence that he had been offered a bribe. His opinion influenced the commissioner of the NFL, Bert Bell, to ban Hapes from the championship game but allow Filchock to play. Amid the furore Filchock gave a gritty performance; despite a broken nose, he threw two touchdowns in a 2414 losing cause. He also was intercepted an amazing six times, consistent with his season-long performance of leading the league with twenty-five interceptions. This led some cynical fans to speculate that Filchock may have tried to lay down in games during the regular season. After the game the commissioner suspended both players until the legal proceedings were completed. According to New York law at that time, attempting to bribe a professional athlete was a felony punishable by one to five years imprisonment and a fine of up to $10,000.

It turned out that Paris was the go-between for three higher-placed gamblers: David Krakauer, Jerome Zarowitz, and Harvey Stemmer. Stemmer was serving time in jail for a previous sports fix while arranging this operation. He was under minimal supervision and was somehow able to exploit the penal system to continue his scheme. Harvey Stemmer was a likeable man, friendly, extroverted, and unassuming. He was also my next-door neighbour when I was growing up in Brooklyn.

Before he got involved with gambling and rigging games, he was a blue-collar worker who toiled hard for modest wages to support his family. The Stemmers were a sports-oriented family, and his son was amazingly well informed about sports statistics at a very early age. Stemmer played with my father in a weekly low-stakes pinochle game, where he was accepted as a regular guy. Gradually he became affiliated with gamblers, and his finances improved. The Stemmers were the first family in our apartment building to own a television set. Somehow the neighbours guessed what was going on. Our apartments were next to each other and shared a fire escape. At one point Stemmer was worried that his telephone was being tapped, and he offered to pay our rent if he could hook up an additional line between our apartments. My mother, afraid of getting mixed up in illegal behaviour, politely declined. Everyone knew why Stemmer was imprisoned after his first sports fix. A year later, when the Daily News headline screamed something like Mystery Harvey Sought in Giants Football Fix, we all knew whom they were looking for.

Appendix B: Raw data

			First test		Final test	
Participant	Feedback	Delay	Score	Average time	Score	Average time
1	no	short	9	5.43	8	4.43
2	no	short	13	6.86	13	4.92
3	no	short	16	9.44	16	4.65
4	no	short	12	5.11	11	4.13
5	no	short	9	9.4	10	5.57
6	no	short	13	7.57	11	6.17
7	no	short	11	7.15	10	6.11
8	no	short	16	4.87	15	3.45
9	no	short	12	9.03	11	6.1
10	no	short	13	7.36	12	4.95
11	yes	short	5	9	11	5.11
12	yes	short	11	6.6	11	6.22
13	yes	short	14	6.65	14	5.28
14	yes	short	11	5.08	15	5.4
15	yes	short	5	4.98	11	6.37
16	yes	short	14	5.48	20	5.67
17	yes	short	12	7.22	15	6.54
18	yes	short	6	6.2	11	5.75
19	yes	short	9	7.95	14	5.83
20	yes	short	11	9.09	14	6.02
21	no	long	10	11.06	10	10.78
22	no	long	12	7.82	12	7.58
23	no	long	15	12.39	16	6.54
24	no	long	9	6.26	10	5.37
25	no	long	11	6.57	13	5.44
26	no	long	6	10.77	6	9.87
27	no	long	8	8.28	9	6.15
28	no	long	13	6.03	14	5.02
29	no	long	9	9.91	10	7.53
30	no	long	13	8.47	12	7.32
31	yes	long	12	5.56	16	4.32
32	yes	long	13	5.81	18	4.56
33	yes	long	13	5.71	19	5.28
34	yes	long	9	5.92	18	4.7
35	yes	long	11	7.32	17	6.46
36	yes	long	11	6.43	16	5.09
37	yes	long	14	6.81	13	5.85
38	yes	long	9	7.28	13	8.52
39	yes	long	10	6.02	14	6.39
40	yes	long	9	7.54	14	6.39