Running head: FEEDBACK AND SPACING EFFECTS

Learning from Feedback: Spacing and the Delay-Retention Effect

Troy A. Smith and Daniel R. Kimball

Department of Psychology

University of Oklahoma

Published in JEP:LMC (http://www.apa.org/pubs/journals/xlm/). This article may not exactly replicate the final version published in the APA journal. It is not the copy of record. The published article should be cited as:

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(1)*, 80-95. doi: 10.1037/a0017407.

Abstract

Most modern research on the effects of feedback during learning assumes that feedback is an error correction mechanism. Recent studies of feedback timing effects suggest that feedback might also strengthen initially correct responses. In an experiment using cued recall of trivia facts, we directly tested several theories of feedback timing effects and also examined the effects of restudy and retest trials following immediate and delayed feedback. Results were not consistent with theories assuming that the only function of feedback is to correct initial errors but instead supported a theoretical account assuming that delaying feedback strengthens initially correct responses due to the spacing of encoding opportunities: Delaying feedback increased the probability of correct response perseveration on the final retention test but had minimal effects on error correction or error perseveration probabilities. In a second experiment, the effects of varying the lags between study, test, and feedback trials during learning provided further support for the spacing hypothesis.

[Word count: 155]

Keywords: spacing effect, feedback timing, delay-retention effect, lag effect

Learning from feedback: Spacing and the delay-retention effect

Recently there has been a resurgence of interest in applying the principles of cognitive psychology to education, and, in particular, examining the role of tests as learning opportunities (for a review, see Roediger & Karpicke, 2006). It has long been known that testing promotes long-term retention, and it is generally agreed that it does so by strengthening the retrieved correct responses (e.g., Bjork & Bjork, 1992; Dempster, 1996; Karpicke & Roediger, 2007). Similarly, providing feedback after a test has consistently been shown to further enhance long-term retention (for reviews, see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Mory, 2004). However, researchers have disagreed as to the role of feedback in the processes underlying learning of factual information, some claiming that it strengthens correct responses (e.g., Skinner, 1953; Butler, Karpicke, & Roediger, 2008), others that it corrects initial errors in learning (e.g., Kulhavy, 1977; Pashler, Cepeda, Wixted, & Rohrer, 2005). Of course, it is likely that feedback actually plays both of these roles, but the extent to which it does one or the other in particular circumstances is not well understood (Butler & Roediger, 2008).

To explore these issues, we focus on a particular aspect of feedback—its timing—in order to further an understanding of the relative contributions of correct-response strengthening and error correction in improving post-feedback performance. In a typical feedback timing experiment, participants study factual information such as trivia questions or an instructional text passage and then take an initial test over the material. Participants are given feedback in the form of the correct answers, either immediately after the test or after a filled delay; and then after a designated retention interval, they are given a retention test. The effect of feedback timing is assessed by comparing retention test performance for the immediate and delayed conditions. In this paper, we report the results of two feedback timing experiments that are designed to test three different theories that have been offered previously as explanations of feedback timing effects.

Previous Feedback Timing Studies

Prior to the mid-1960's it was generally believed that the principles of classical and operant conditioning that guide animal learning also guide human learning, including learning of factual information. Consistent with this belief, early studies of human learning using stimulus-response learning tasks showed that immediate feedback led to better learning than did delayed feedback (see Renner, 1964 for a review), and most researchers assumed these results would generalize to other learning tasks such as those commonly used in educational settings (Pressey, 1963). However, a series of studies conducted in the early 1960's that were designed to test this assumption (Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962) found just the opposite: Delaying feedback after an initial test generally resulted in an improvement in performance on a subsequent, delayed test.

This finding is now known as the *delay-retention effect* (Kulhavy & Anderson, 1972), and it has been widely replicated. In the experiments by Brackbill and colleagues elementary school children learned two-choice discriminations better when feedback was delayed a few seconds than when there was no delay. Over the next two decades, the delay retention effect was observed in a number of studies that used multiple-choice tests and showed the effect with more complex stimuli such as general factual knowledge (Sturges & Crawford, 1964); with a more diverse population, including junior high school students (More, 1969), high school students (Kulhavy & Anderson, 1972; Surber & Anderson, 1975), and college undergraduates (Sassenrath & Yonge, 1968, 1969); and, perhaps most importantly, with feedback delays ranging from 15s (Rankin & Trepper, 1978) to as long as four days (More, 1969).

The delay-retention effect appears to be fairly robust when using multiple-choice tests in controlled settings. In a meta-analysis of feedback timing studies, Kulik and Kulik (1988) found an advantage for delayed feedback relative to immediate feedback in all but one of the 8 feedback timing studies in their analysis that used multiple-choice tests to assess learning after retention intervals of 5-7 days. However, there are questions as to the generality of this finding. When Kulik and Kulik examined studies that were more applied in nature, they concluded that immediate feedback was preferable. In our own review of 39 different studies of feedback timing effects on long-term retention, we found 16 studies that showed a significant advantage for delayed feedback, 12 that showed a significant advantage for immediate feedback, and 11 that failed to find any significant effect of feedback timing (Smith, 2007). One of the goals of the experiments we report in this article was to explore potential reasons for the discordant pattern of findings in previous studies, a point to which we return in the General Discussion.

Theories of Feedback Timing Effects

Given that there is a certain amount of confusion in the literature as to the empirical effects of feedback timing on retention, it is not surprising that there is also some disagreement about the theoretical explanations of feedback timing effects. This disagreement revolves around the central question: What is the function of feedback, i.e., how does it improve memory performance? In the next section, we discuss three theories of feedback timing. The first two—learning theory and the spacing hypothesis—are based on the assumption that feedback functions to strengthen correct responses, although they differ as to how feedback timing affects that strengthening. The third theory—the interference perseveration hypothesis—is based on the assumption that feedback functions to correct errors.

Learning Theory

The earliest theoretical accounts of the effect of feedback were grounded in learning theory, a combination of the principles of classical and operant conditioning, according to which feedback is assumed to function as a reinforcer of correct responses. These accounts assumed that the principles of conditioning that guide animal learning also guide human learning, including learning of verbal and factual information (Pressey, 1963; Skinner, 1953). A central tenet of these conditioning theories is that the effect of a reinforcer is optimal when it is provided in close temporal proximity to the response that is to be reinforced and that the effectiveness of reinforcement is reduced as the temporal proximity decreases (Hull, 1952; Saltzman, 1951; Skinner, 1953). Thus, learning theory predicts that immediate feedback should reinforce correct responses and thereby increase retention, and that delaying feedback by more than a few seconds should eliminate these effects (Renner, 1964).

Early studies of feedback timing effects in educational settings provided evidence that also seemed to support this prediction (e.g., Angell, 1943). However, because learning theory does not appear to be compatible with the later studies that showed a delay-retention effect, it was rejected by most feedback researchers during the 1960's and then largely ignored thereafter. As a byproduct of this rejection, researchers also largely ignored the assumption underlying learning theory that correct response strengthening can contribute to feedback effects in general and to feedback timing effects in particular (Mory, 2004).

The Spacing Hypothesis

Recently, though, Smith, Kimball, and Mann (2007); Butler, Karpicke and Roediger (2007); and Pashler, Rohrer, Cepeda, and Carpenter (2007) independently proposed an alternative account of feedback timing effects that is based on this previously ignored assumption that feedback functions to strengthen correct responses. Rather than viewing feedback through the lens of conditioning theory—in which feedback is a mechanism for reinforcing responses to stimuli—this new account views feedback through the lens of memory theory, according to which feedback trials serve as additional encoding opportunities. From this viewpoint, and contrary to the prediction made by learning theory, delayed feedback trials should tend to strengthen initially correct responses more than should immediate feedback trials because of differences in the distribution of the encoding opportunities. We refer to this account as the *spacing hypothesis*.

The spacing hypothesis is based on two observations. First, delayed feedback inherently acts as a spaced restudy opportunity for the items that were initially answered correctly, whereas immediate feedback acts as a massed study opportunity for these items (Butler et al., 2007; Pashler et al., 2007; Smith et al., 2007). Second, distributed learning episodes lead to better learning than do massed learning episodes as measured by performance on a delayed retention test (see, e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1996). Thus, the spacing hypothesis predicts the delay-retention effect because delaying feedback increases the probability that correct responses will perseverate to the retention test due to a spacing effect, and—assuming enough responses are correct on the initial test—delaying feedback thereby tends to increase long-term retention overall.

If feedback timing effects are a type of spacing effect as proposed by the spacing hypothesis, then this implies that feedback timing effects should be subject to the same constraints as other spacing effects and that they can be explained by the same underlying mechanisms (for reviews, see Cepeda et al., 2006; Dempster, 1996; or Hintzman, 1974). Additionally, classifying feedback timing effects along with other types of spacing effects such as repeated testing could provide additional constraints that might help evaluate the various mechanisms that have been proposed as explanations of spacing effects. We address this point in more detail in the General Discussion.

The Interference Perseveration Hypothesis

Of course, whether or not feedback strengthens correct responses, it is reasonable to assume that it could also allow learners opportunities to correct mistakes and thereby improve their performance on later tests. Guthrie (1971) proposed that during the learning of new information the primary function of feedback is to correct initial errors in learning, and this assumption has guided most feedback research conducted since then (Pashler et al., 2005). Notwithstanding the compatibility of error correction and correct-response strengthening, explanations of feedback timing effects developed under the Guthrian assumption have generally assumed that feedback has no effect on initially correct responses (Surber & Anderson, 1975) or, at most, confirms these responses so that the learner maintains them in memory instead of discarding them (Kulhavy, 1977).

Following Guthrie's (1971) lead, Kulhavy and Anderson (1972) offered what has become the leading theoretical explanation for feedback timing effects, the *interference perseveration hypothesis*. The interference perseveration hypothesis is based on the similarity between the feedback timing paradigm and the AB-AD paradigm that is commonly used to study interference effects. In Kulhavy and Anderson's (1972) analysis, the initial test trial in the feedback timing paradigm is analogous to the initial AB trial, with the test question serving as the A cue and the participant's response to the test question serving as the B target. If the participant's initial response is correct, then the feedback trial becomes another AB trial. This second AB trial might strengthen the response tendency, but because there is no interference, this theory assumes that there should be no effect of delay on this strengthening tendency (Kulhavy, 1977; Kulhavy & Anderson). However, if the participant's initial B response is incorrect, then the correct response serves as the D target, the feedback trial becomes an AD trial and the incorrect B response should proactively interfere with the learning of the association between the question and the correct answer.

Kulhavy and Anderson (1972) applied this idea to the effects of feedback timing, positing that:

> learners forget their incorrect responses over the delay interval, and thus there is less interference with learning the correct answers from the feedback. The subjects who receive immediate feedback, on the other hand, suffer from proactive interference because of the incorrect response to which they have committed themselves. (pg. 506)

Thus, the interference perseveration hypothesis assumes that a) the function of feedback is to correct errors, b) this error correction function is impaired when feedback is given immediately, and c) introducing a delay allows the error correction function to operate more effectively. Because the interference perseveration hypothesis has seemed to be an elegant explanation of the delay-retention effect that is based on well-established memory theory, it has been a leading explanation of feedback timing effects for over 35 years (Butler & Roediger, 2008; Clariana, Wagner, & Murphy, 2000; Kulik & Kulik, 1988; Mory, 2004; Swindell & Walls, 1993).

Summary

The three theories discussed above make different predictions about how varying the timing of feedback should affect long-term retention and whether the effect of timing is driven by correct response perseveration or error correction. Learning theory predicts that immediate feedback should lead to better retention because it more effectively reinforces correct responses.

The spacing hypothesis predicts that delayed feedback should increase correct response perseveration due to the greater distribution of processing opportunities. The interference perseveration hypothesis also predicts a delay-retention effect but predicts that the effect will be due to a decrease in error perseveration and a corresponding increase in error correction with a delay.

Overview of Present Experiments

The experiments we report are designed to test the predictions of the feedback timing theories discussed above. The experiments are designed to use stimuli and procedures similar to those used in classroom learning or computer aided instruction while also maintaining the degree of experimental control afforded within a laboratory setting. To increase ecological validity, the stimuli were trivia facts that participants were likely to find interesting but were unlikely to know prior to beginning the experiment, and we had participants study the stimulus materials prior to the initial test—a step that some feedback timing studies have not included (e.g., Kulhavy & Anderson, 1972). Additionally, we used inter-trial intervals on a scale of interest to educators and theoreticians alike. However, to maintain control over stimulus exposure, we strictly controlled the amount of time allowed for each study, test, and feedback trial.

In our experimental designs, we also addressed a number of methodological issues that may be responsible for much of the confusion in the feedback timing literature. First, almost every reported study examining the effects of feedback timing on long-term retention has used multiple-choice tests (usually four alternative forced choice, 4AFC), despite the fact that such tests are sub-optimal in evaluating interference theories (Ausubel, Stager, & Gaite, 1969; Dean, Garabedian, & Yekovich, 1983; Kane & Anderson, 1978; Postman, Stark, & Burns, 1974). To provide a clearer test of the interference perseveration theory than is possible with multiple-choice tests, we used cued recall tests during both the initial learning phase and the retention test phase. Second, in order to better discriminate among the theories, we conducted a fine-grained analysis of the effects of feedback timing on the conditional probabilities that initial errors either were corrected or perseverated to the retention test, and that initially correct responses perseverated. Finally, to provide appropriate tests of feedback timing theories discussed previously, we defined immediate feedback as providing the correct answer immediately after the test question (i.e., item-by-item feedback) and delayed feedback as providing the answer after a delay of at least several minutes. We return to these methodological points in the General Discussion.

Experiment 1

In Experiment 1, we sought to investigate the effect of immediate and delayed feedback on learning of factual material when cued recall tests are used. In addition to the basic manipulation of feedback timing, we also manipulated two other variables that we predicted would be impacted by feedback timing. The first of these was the amount of time between the initial study trial and the initial test trial (i.e., the initial study-test lag). Each of the theories we are testing predicts that feedback timing should have different effects on initially correct responses than on initially incorrect responses. Increasing the initial study-test lag should lead to more errors and retrieval failures on the initial test, thereby potentially modulating the effects of feedback timing. To test for this potential interaction, we included a pair of conditions in which we doubled the duration of the initial study-test lag.

The other added variable was a manipulation of the occurrence and type of post-feedback processing opportunities. Because many previous feedback timing studies—especially those that have focused on applications to classroom learning (e.g., Brosvic, Epstein, Cook, & Dihoff, 2005)—have not controlled post-feedback processing opportunities, we wanted to test whether post-feedback processing opportunities modulate feedback timing effects. Therefore, we include conditions in which participants were given additional delayed study or test trials after receiving feedback. This manipulation also allowed us to test two predictions of the spacing hypothesis: First, because these post-feedback processing opportunities provide additional distributed practice, they should increase correct response perseveration rates relative to when items do not receive such opportunities. Second, as a consequence of the diminishing returns of spacing as the number of distributed learning trials increases (Cull, 2000), items receiving delayed feedback should benefit less from post-feedback processing than items receiving immediate feedback, potentially negating any benefit of delayed feedback that would otherwise occur.

Method

Participants

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

Design

Experiment 1 used a within-subjects design with eight conditions. The design comprised two overlapping factorial manipulations as shown in Figure 1. The first portion of the design formed a 2 × 2 factorial manipulation of feedback timing (immediate vs. delayed) and initial study-test lag (long vs. short). The second portion of the design formed a 2 × 3 factorial manipulation of feedback timing (immediate vs. delayed) and post-feedback processing opportunities during the learning phase (none, restudy, or retest). For convenience, we refer to the 2 × 2 portion of the design as the *timing × lag* conditions and the 2 × 3 portion of the design as the *timing × post-feedback processing* conditions.

Materials

A total of 76 trivia facts were selected from a database of trivia facts that were normed for base knowledge rate and interestingness in a separate study conducted at the University of Texas at Arlington. The trivia facts in the database were drawn from a variety of general knowledge domains. Each trivia fact was in the form of a question and a short answer of one to two words or numbers. For example, "The first zoo in the United States was opened in which city?" (Philadelphia). For this experiment, facts were selected that were rated as highly interesting by one set of respondents in the norming study, but for which none of the respondents in a different set knew the correct answer beforehand. Twelve facts were used as buffer items to control for primacy effects and to help maintain spacing of the trials, and the other 64 facts were randomly assigned to the experimental conditions such that there were 8 facts in each condition. The assignment of facts to conditions and the presentation order of the facts within each condition were randomly determined anew for each participant.

Procedure

The experiment was conducted in two sessions: a 60 min learning session on the first day and a 30 min retention test session one week later at the same time of day. Informed consent was obtained before the learning session began and again before the test session began. Participants were run individually using a custom computer program to present the trivia facts and to record the participants' responses.

A scheduling algorithm similar to the one used by Pashler, Zarow, and Triplett (2003) was used to schedule the study, test, feedback, restudy, and retest trials for each trivia fact during the learning session. Figure 1 shows the presentation schedule for each condition during the learning session. At the beginning of the learning session, the complete set of trivia questions and answers were presented, one fact at a time for 10s each. Following these initial study trials, a series of interleaved test, feedback, restudy, and retest trials began. Pursuant to the scheduling algorithm, the trials for any particular trivia question were interleaved amongst trials for other trivia questions in such a way as to implement the appropriate intervals between the trials for each trivia question assigned to a particular condition. The first test trial for each question occurred approximately 8-10 min after the initial study trial in the short study-test lag conditions or 16-20 min after the initial study trial in the long study-test lag conditions. Feedback trials occurred immediately after the initial test trial in the immediate feedback conditions or approximately 8 min after the initial test trial in the delayed feedback conditions. In the restudy and retest conditions an additional study or test trial, respectively, occurred approximately 8 min after the feedback trial.

On the study, feedback, and restudy trials, participants were shown a question, the correct answer, and a response text box; on test and retest trials, participants were shown the question and a response text box. Participants were given 10s to type the correct answer into the text box for all trials, including the study, feedback, and restudy trials, before the computer automatically moved on to the next trial; this was done to ensure that all of the learning task trials were as equivalent as possible. Participants were also asked to judge how confident they were that their answer was correct using a 7-point Likert scale, with 1 representing “Not at all confident” and 7 representing “Very confident”[1].

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

Results

Performance on the retention test was analyzed using separate repeated measures ANOVAs and sets of planned comparisons for each of the two overlapping factorial manipulations. Additionally, conditional analyses were conducted in order to examine differences in the probabilities of error correction, error perseveration, and correct response perseveration between immediate and delayed feedback (i.e, across the rows in Figure 1). Alpha was set to .05 for all inferential analyses.

Scoring

Participants' responses on each trial were scored using both strict and lenient scoring criteria. For strict scoring, only responses that were identical to the correct answer (ignoring upper- and lower-case) were counted as correct. For lenient scoring, responses that semantically matched the correct answer—such as phonetically spelled responses and incomplete responses that could only have been reasonably completed as the correct answer—were counted as correct Planned analyses were conducted using both sets of scoring criteria, and the basic pattern of results was similar for both criteria. Because the lenient scoring criteria more accurately reflect performance as it would be assessed in educational settings, we report only the lenient scoring results.

Effects of Feedback Timing and Initial Study-Test Lag

As a manipulation check, we compared the scores on the initial test for the short and long initial study-test lag conditions. As expected, initial test performance was better in the short lag conditions ($M = .70$, SE = .016) than in the long lag conditions ($M = .68$, SE = .019), $F(1, 102) = 3.98$, $MSE = 0.013$, $p = .05$. However, this effect was somewhat smaller than we had hoped, $\eta^2 = .038$.

Table 1 shows the retention test scores for the timing × lag conditions. A 2 (feedback timing: immediate; delayed) × 2 (initial study-test lag: short; long) repeated measures ANOVA on these conditions revealed a significant main effect of feedback timing on retention test scores, $F(1,102) = 12.76$, $MSE = 0.030$, $p < .01$, partial $\eta^2 = .111$, with delayed feedback leading to better retention test performance than did immediate feedback. There was no significant main effect of initial study-test lag, $F < 1$; but there was a significant interaction of feedback timing and study-test lag, $F(1,102) = 11.10$, $MSE = 0.028$, $p < .01$, partial $\eta^2 = .098$. As can be seen in Table 1, this interaction reflected a delay retention effect when the study-test lag was shorter, $t(102) = 5.05$, $SE = .023$, $p < .01$, and the disappearance of that effect when the study-test lag was longer, $t(102) = 0.25$, $SE = .025$, $p > .10$. We offer an explanation for this pattern in the Discussion section.

Effects of Feedback Timing and Post-Feedback Processing

Table 2 shows the retention test scores for the timing x post-feedback processing conditions. A 2 (feedback timing: immediate; delayed) × 3 (post-feedback processing: none; restudy; retest) repeated measures ANOVA revealed a significant main effect of feedback timing on retention test scores, $F(1,102) = 16.05$, $MSE = 0.020$, $p < .01$, partial $\eta^2 = .136$, such that delaying feedback improved retention overall compared to providing immediate feedback. There was also a main effect of post-feedback processing, *F*(2,204) = 14.14, *MSE* = 0.028, $p < .01$, partial $\eta^2 = .122$, such that compared to no post-feedback processing, retention test scores improved with restudy of the material following feedback, *F*(1,102) = 10.00, *MSE* = 0.117, $p < .01$ partial $\eta^2 = .089$, and scores following a retest were even better than scores following restudy, *F*(1,102) = 4.96, *MSE* = 0.089, $p = .03$, partial $\eta^2 = .046$.

These main effects were qualified by a significant interaction of feedback timing and post-feedback processing, *F*(2,204) = 7.80, *MSE* = 0.025, $p < .01$, partial $\eta^2 = .071$, as predicted by the spacing hypothesis. In the absence of any post-feedback processing, delaying feedback improved retention test performance relative to providing immediate feedback, *t*(102) = 5.05, *SE* = .023, $p < .01$. However, this effect was eliminated when the feedback trial was followed by a delayed restudy trial, *t*(102) = 0.25, *SE* = .019, $p > .10$, or retest trial, *t*(102) = 0.79, *SE* = .022, $p > .10$. As can be seen from the means in Table 2, the interaction was primarily due to the similarity in the simple effects of feedback delay and post-feedback processing: Relative to the condition involving immediate feedback without post-feedback processing, providing a delayed restudy or retest trial after the immediate feedback trial increased retention test scores, *t*(102) = 4.52, *SE* = .024, $p < .01$ and *t*(102) = 5.32, *SE* = .026, $p < .01$, respectively, just as delaying feedback did.

Conditional Analyses

In order to test the relative contributions of correct response perseveration, error correction, and error perseveration to feedback timing effects, we examined the response probabilities for the retention test conditionalized on participants' responses on the initial test. We started by aggregating responses from all participants and calculating the mean conditional response probabilities for each of the eight experimental conditions (see Table 3) in the same manner as Butler et al. (2007). As Butler et al. point out, aggregating data across all participants avoids the problems inherent in dealing with missing data in the conditional cells for which some individual participants did not generate observations (e.g., repetition of an initial incorrect response, $I_{old} \mid I_{initial}$)[2].

However, because aggregating data across participants also introduces violations of the assumptions that underlie traditional inferential statistics, standard techniques for conducting null-hypothesis tests on aggregated data do not exist (Schafer & Graham, 2002). Therefore, we supplemented the aggregated conditional analyses with a set of bootstrap analyses designed to obtain reliable estimates of the null-hypothesis distributions for contrasts involving correct response perseveration, error correction, and error perseveration probabilities[3]. We then compared these distributions to the contrast values calculated from the observed (aggregated) mean probabilities in order to conduct quantitative tests for the hypotheses of interest. A detailed description of the bootstrap procedure is provided in the Appendix

Correct Response Perseveration. An examination of the conditional response probabilities for items that were answered correctly on the initial test shows that, in general, relative to providing immediate feedback, delaying feedback increased the probability of repeating an initially correct response. As can be seen in Figure 2 (left panel in each row), this effect was significant in the no re-processing (short lag) condition, $\Delta M = .15$, bootstrapped $p < .001$, as predicted by the spacing hypothesis but counter to predictions from learning theory. However, the effect was substantially attenuated when the feedback trial was followed by a delayed restudy trial, $\Delta M = .03$, bootstrapped $p = .11$, or a delayed retest trial, $\Delta M = .05$, bootstrapped $p < .02$. As Table 3 shows, this attenuation in the effect of feedback delay when there was a post-feedback restudy or retest trial was due primarily to an increase in the correct response perseveration probabilities for the immediate feedback condition when there was some re-processing as compared to when there was not, consistent with the spacing hypothesis. The effect of feedback timing on correct response perseveration was not significant for the long study-test lag condition, bootstrapped $p > .10$.

Error Correction and Perseveration. Figure 2 (middle and rightmost panels in each row) clearly shows that feedback timing did not have a systematic effect on error correction or error perseveration conditional response probabilities (see also Table 3), contrary to predictions from the interference perseveration hypothesis. In fact, the only condition for which there was an effect of feedback timing was when the feedback trial was followed by a delayed retest trial. In that condition, delaying feedback led to a significant increase in perseverated errors, $\Delta M = .10$, bootstrapped $p = .036$, and a corresponding decrease in error correction, $\Delta M = -.15$, bootstrapped $p = .043$; however, the direction of this effect is opposite that predicted by the interference perseveration hypothesis. In the other conditions, there were no significant effects of feedback timing on items that were answered incorrectly on the initial test, all *p*s > .10.

Discussion

In Experiment 1, we observed the delay-retention effect with cued recall tests: In general, delayed feedback led to better performance on a test after a 7-day retention interval than did immediate feedback. This finding is not consistent with predictions from learning theory but is consistent with predictions from the spacing hypothesis and the interference perseveration hypothesis. Analyses of conditional response probabilities showed that the difference in long-term retention was primarily caused by the increased probability of remembering an initially correct response when feedback was delayed, as predicted by the spacing hypothesis. Contrary to predictions from the interference perseveration hypothesis, there was no evidence that error correction or error perseveration probabilities differed systematically as a function of feedback timing.

Additional evidence supporting the spacing hypothesis is provided by secondary results in Experiment 1. First, the effect of feedback delay was moderated by additional processing of the stimuli after the feedback trial, with restudy and retest trials attenuating or eliminating the effect of feedback delay relative to conditions that did not include additional processing trials. Moreover, this attenuation was attributable to an increase in correct response perseveration for the immediate feedback condition relative to the no re-presentation condition, as predicted by the spacing hypothesis. Second, when spacing was equated by adding a delayed restudy trial to the immediate feedback condition, the rate of correct response perseveration increased in the same manner that it increased when feedback was delayed, exactly as predicted by the spacing hypothesis. Third, the delay retention effect observed with a shorter lag between study and initial test disappeared when that lag was lengthened. Conditional analyses revealed that correct response perseveration was affected by feedback timing only at the shorter lag and not at the longer lag, and that feedback timing did not affect error correction or perseveration at either lag. These findings are consistent with typical effects of lag on correct responses, as predicted by the spacing hypothesis: When lag is lengthened by a moderate amount, as in this case, correct responses are strengthened to a greater degree. We discuss lag effects in more detail in connection with Experiment 2.

In Experiment 1 the timing of all trials was strictly controlled in order to avoid potential confounds between feedback timing and the amount of time participants spent studying feedback. This ensured that the results could be clearly interpreted in terms of the three theories we are testing, but it raises the question of how well these results might generalize to real-world learning conditions in which learners control the amount of time they spend studying. To address this issue, we conducted an experiment, not reported here, that was identical to Experiment 1 except that we held study time constant only for the initial study trials, and allowed participants to self-pace the other trials (test, feedback, restudy, and retest trials). The results from this experiment replicated all our critical findings. This suggests that theses results are robust and that the results generalize to somewhat more real-world learning conditions.

Although Experiment 1 showed consistent results for the effects of feedback timing and post-feedback presentation, the effects of initial study-test lag were not as clear. There was a small but reliable effect of initial study-test lag on initial test performance, but this did not transfer to the retention test one week later. Of course, a possible reason for these results is that the difference between an 8-10 min initial study-test lag and a 16-20 min lag is relatively small relative to the 7-day retention interval. In the next experiment we explored the effects of varying the time lag between study, test, and feedback trials over a wider range.

Experiment 2

If, as suggested by Experiment 1, feedback timing effects are a type of distributed practice effect, then manipulations of the lag between study, test, and feedback trials ought to show the same effects of lag that have been observed in distributed practice studies. Experiment 2 was designed to examine two of these effects: a) a non-monotonic effect of inter-trial lag on final recall; and b) a benefit in final recall for an expanding practice schedule over a uniform or contracting schedule.

The non-monotonic relationship between retention and inter-trial lag in cued recall was one of the earliest identified patterns in the distributed practice literature. In two experiments using the continuous paired-associate paradigm, Peterson, Wampler, Kirkpatrick, and Saltzman

(1963) found that as the spacing interval (lag) between presentations increased, retention showed an inverted-U pattern (see also Balota, Duchek, & Paullin, 1989; and Glenberg, 1976). A recent meta-analysis conducted by Cepeda et al. (2006) showed that the nonmonotonic shape of the lag effect curve generalizes across a variety of stimuli and across a wide range of lags and retention intervals.

Figure 3 presents idealized versions of the lag effect curves for short and long retention intervals (left and right panels, respectively). As shown by the left-hand portion of the two curves in Figure 3, if the lag is not too long relative to the retention interval, increases in lag lead to increases in performance on a retention test. For any given retention interval, there is a lag that maximizes performance (the peaks, Points D and E', of the curves in Figure 3); increases in lag beyond this optimal value result in decreases in performance (the right-hand portion of the curves in Figure 3). The optimal lag tends to increase non-linearly as the retention interval increases (compare the locations of Points D and E' in Figure 3), with the optimal lag for long retention intervals (such as we use in this experiment) being between 5% and 20% of the retention interval (Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler, 2009; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008).

Because the spacing hypothesis of feedback timing effects assumes that study trials, successful test trials, and feedback trials all strengthen correct responses, it predicts that the lag effect pattern described above should hold for variations in study-test lag and test-feedback lag. Because items that are answered correctly on the initial test receive three spaced processing opportunities whereas items not answered correctly effectively receive two spaced processing opportunities, this pattern should be particularly prominent for items that were answered correctly on the initial test. We directly test these predictions in Experiment 2 by comparing performance after retention intervals from 5 to 7 days for conditions in which the total lag across study, test, and feedback trials ranged from approximately 16 minutes (< 1% of the retention interval) to 24 hours (15-20% of the retention interval) to 48 hours (~40% of the retention interval). If the spacing hypothesis is correct, and the temporal distribution of successful test trials and feedback trials operates much like the temporal distribution of study trials, we should observe a curvilinear relationship between total lag and recall, with optimal retention in the 24-hour lag condition.

We also examined the impact of the relative lags between the trials for a particular item by comparing recall in conditions with lags between the trials for a particular item that were uniformly sized, increasing in size (i.e., an expanding practice schedule), or decreasing in size (i.e., a contracting practice schedule). Landauer and Bjork (1978) first reported the superiority in recall that results from an expanding retrieval practice schedule as compared to uniform or contracting schedules, and this finding was later replicated and extended by Cull, Shaughnessy, and Zechmeister (1996; Experiments 1-4). In these previous studies, all the trials in the practice schedule were test trials without feedback. We sought to extend these findings to practice schedules that include a delayed feedback trial in addition to a test trial.

Method

Participants

Participants were 81 undergraduate students enrolled in psychology courses at the University of Oklahoma who participated for partial course credit. Data from participants who failed to complete all four experimental sessions ($n = 7$), and data that were not properly stored due to experimenter error (n = 1), were excluded from further analysis.

Design

For Experiment 2, we used a within-subjects design with 10 conditions, as shown in Table 4. The learning phase of the experiment took place over three days (Days 1, 2, and 3) and the retention test was administered on Day 8. The key differences among the 10 conditions are in the amount of time between the study trial and the initial test trial (study-test lag), the amount of time between the initial test trial and the feedback trial (test-feedback lag), and the amount of time between the feedback trial and the retention test trial (retention interval). The 10 conditions comprise all combinations of study-test lag (0, 1, or 2 days), test-feedback lag (0, 1, or 2 days), and post-feedback retention interval (5, 6, or 7 days) that were possible given the timing of the four sessions. Comparisons across different sets of these conditions allowed us to examine the effects on final recall of the total lag between study and feedback trials (0, 1, or 2 days) and of different practice schedules (with uniform, expanding, or contracting lags), as we describe in more detail below. These comparisons were greatly simplified due to the inconsequential effects of post-feedback retention interval, which allowed us to collapse across that variable.

To examine the effects of total study-feedback lag on final recall, we compared recall across three sets of conditions, which are identified in the second column of Table 4. In the first set, the study, test, and feedback trials all occurred on the same day (the *0-day lag* condition, including Conditions 1, 2, and 3). In the second set, the study and feedback trials occurred 1 day apart (the *1-day lag* condition, including Conditions 4, 5, 6, and 7). In the third set, the study and feedback trials occurred 2 days apart (the *2-day lag* conditions, including Conditions 8, 9, and 10).

To examine the effect of different practice schedules, we classified each of the experimental conditions as having either a uniform, expanding, or contracting schedule. These classifications are set forth in the third column of Table 4. In the *uniform* schedule condition (including Conditions 1, 2, 3, and 8), the study-test lag was equal to the test-feedback lag; in the *expanding* schedule condition (including Conditions 4, 5, and 9), the test-feedback lag was longer than the study-test lag; and in the *contracting* schedule condition (including Conditions 6, 7, and 10), the test-feedback lag was shorter than the study-test lag. This classification allowed us to conduct an overall analysis that collapsed across all conditions within each type of schedule, as well as an analysis that compared only those conditions having the same study-feedback lag and retention interval (i.e., Conditions 8, 9, and 10).

Materials

A total of 96 trivia facts were selected from the same database as those used in Experiment 1, using similar criteria. Six facts were used as buffers to control for primacy and recency, and the remaining facts were randomly assigned to the 10 experimental conditions such that there were 9 facts in each condition. The assignment of facts to conditions and the presentation order of the facts within each condition were randomly determined anew for each participant. As in Experiment 1, a scheduling algorithm implemented the inter-trial intervals for each item in a particular condition by interleaving those trials with the trials for other items.

Procedure

The experiment was conducted in four sessions: three 30-min learning sessions on Days 1, 2, and 3; and a 30-min retention test session on Day 8. All three learning sessions and the retention test session were scheduled at approximately the same time of day: Due to day-to-day variations in our participant's schedules, we allowed the appointment times for each day to be within the same four hour window. As in Experiment 1, participants were run individually using a custom computer program to present the trivia facts and to record the participants' responses.

Across the three learning sessions, each fact was presented for one study trial, one initial test trial, and one feedback trial as shown in Table 4. Participants were asked to type the correct answer on all trials in the learning sessions, including study and feedback trials. During the retention test on Day 8, each of the 96 facts was presented for a retention test trial. Both the initial test trial and the retention test trial were cued recall trials, with the trivia question serving as the cue, and all trials in both sessions lasted 12s each with a blank screen displayed for 500ms between each trial.

Results

Scoring

Three separate raters independently scored the participants' responses for each test trial using the same criteria that were used for lenient scoring in Experiment 1. The three raters agreed on more than 95% of participant responses, and the remaining items were scored based on the agreement of 2 of the 3 raters.

Performance on the Initial Test

As a check to ensure that our study-test lag manipulation effectively induced differential levels of forgetting prior to the presentation of feedback, we examined performance on the initial test as a function of study-test lag. Table 4 lists the mean proportions of correct responses on this initial test for the 10 experimental conditions. A set of planned orthogonal contrasts revealed that, as expected and consistent with the standard forgetting curve, participants scored highest on the initial test when it was on the same day as the study trial (Conditions 1, 2, 3, 4, 5, and 9), $M = .85$, $SE = .011$, $t(72) = 25.48$, $SE = 0.012$, $p < .01$; lower after a 1-day study-test lag (Conditions 6, 7, and 8), $M = .62$, $SE = .016$, $t(72) = 20.11$, $SE = 0.012$, $p < .01$; and lower still after a 2-day study-test lag (Condition 10), $M = .49$, $SE = .018$, $t(72) = 6.36$, $SE = 0.020$, $p < .01$. This pattern of decreasing initial test performance also held as the study-feedback lag increased (see Figure 4).

Retention Test Performance as a Function of Study-Feedback Lag

As predicted by the spacing hypothesis, there was a non-monotonic relationship between inter-trial lag and retention test performance. The lag effect curve shown in Figure 4 for the retention test shows the inverted-U shape that is typical in distributed practice studies. A set of planned within-subjects contrasts confirmed that recall was higher in the 1-day lag than in the 0-day and 2-day lag conditions combined, thus signifying the presence of a significant quadratic trend, $t(72) = 3.47$, $SE = .010$, $p < .01$, and that there was no significant difference in recall for the 0-day and 2-day lag conditions, $t(72) = -1.97$, $SE = 0.016$, $p > .05$.

We also examined the effects of lag conditionalized on participants' initial test responses. As in Experiment 1, these conditional analyses were conducted by aggregating responses in each experimental condition across all participants to calculate means. Confidence intervals and *p* values were calculated for the contrasts of interest using the bootstrap method described in the Appendix. As can be seen in Figure 5, increasing study-feedback lag from 0 days to 1 or 2 days increased both the probability of repetition for initial correct responses ($C_{ret} \mid C_{initial}$), bootstrapped $p < .001$, and the probability of correction for incorrect and blank initial test responses combined ($C_{ret} \mid \text{not } C_{initial}$), bootstrapped $p < .001$. Consistent with the predictions of the spacing hypothesis, this increase was negatively accelerated in both cases, as revealed by the presence of significant quadratic trends, bootstrapped *p*s = .002 and .036, respectively, although the increase for correct response perseveration may have been limited by a ceiling effect.

Retention Test Performance as a Function of Presentation Schedule

To examine the effects of the different types of presentation schedules, we conducted a set of planned within-subject contrasts. Overall, across all 10 conditions, an expanding schedule (M = .78, SE = .016) led to better retention test performance than either a uniform schedule (M = .72, SE = .018), $t(72) = 5.07$, SE = .012, $p < .01$, or a contracting schedule (M = .73, SE = .017), $t(72) = 3.29$, SE = .014, $p < .01$. There was no significant difference between a uniform and a contracting schedule, $p > .10$.

In isolation, these omnibus results would have to be interpreted with caution because there are differences between the expanding, uniform, and contracting groups other than just the practice schedule. For example, on average, there are fewer days between learning episodes in the uniform conditions than in the expanding or contracting conditions. Therefore, we conducted a separate analysis in which we compared only the conditions that had the same two-day lag between study and feedback trials and the same five-day post-feedback retention interval (uniform: Condition 8; expanding: Condition 9; and contracting: Condition 10). The same pattern was revealed in this analysis as in the overall analysis: an expanding retrieval practice schedule was better than a uniform schedule, $t(72) = 2.16$, SE = .023, $p = .03$, or contracting schedule, $t(72) = 3.55$, SE = .023, $p < .01$—consistent with findings reported by Landauer and Bjork (1978)—and there was no significant difference between a uniform and a contracting schedule, $t(72) = 1.38$, SE = .023, $p > .10$.

Discussion

Experiment 2 demonstrated that the feedback timing paradigm is subject to the same type of lag effects as is the distributed practice paradigm. Consistent with predictions from the spacing hypothesis and the work of Cepeda and colleagues (Cepeda et al., 2009; Cepeda et al., 2006; Cepeda et al, 2008), we found that retention test scores and correct response perseveration probabilities increased as the lag between an item's first and last learning-phase trials increased from less than 1% of the retention interval to approximately 20% of the retention interval. However, consistent with results reported in the literature on distributed practice, there were diminishing returns and the beginning of a downward trend with a further increase in lag to approximately 40% of the retention interval.

According to theories of spacing effects, the inverted-U shape of the lag effect curve results from the composition of a forgetting function with a response strengthening function (Cepeda et al., 2009). For example, encoding variability theories assume that forgetting between learning trials occurs as a result of contextual change or drift and that increasing the inter-trial lag increases contextual change between encoding opportunities (Balota et al., 1989; Glenberg, 1976). This increased contextual change has two countervailing effects. First, it causes performance on test trials during the learning session to decrease. Second, for items that were successfully retrieved on the test trial, it increases the probability that an encoded context will be retrieved on a later retention test, thereby increasing the probability of recall for those items (Glenberg, 1976).

Notably, this is precisely the pattern of results obtained in Experiment 2. As the inter-trial lag increased, performance on the initial test decreased (see Figure 4), while the conditional probability of initially correct responses being repeated on the retention test increased (see Figure 5). Because our paradigm included a delayed feedback trial that could serve as a second study trial for the items that were not answered correctly on the initial test, we also observed a lag effect for those items as well.

Although the non-monotonicity in the lag effect curve in Experiment 2 was not as large as that observed in some studies, this is likely due to procedural differences that affected the level of initial learning. For example, in Cepeda et al.'s (2008) study, participants learned 32 trivia facts to a perfect criterion in the initial study session and then had two test cycles with immediate feedback in the second session before taking a final retention test. By contrast, in our experiment, participants had to learn many more trivia facts, and they only had one study trial prior to the initial test.

Experiment 2 also showed that with the multiple processing opportunities that are inherently part of the feedback timing paradigm, an expanding schedule of trials in which there was a short study-test lag and a longer test-feedback lag lead to better performance after a one week retention interval than did a uniform or contracting schedule, consistent with results that have been reported by Landauer and Bjork (1978) and Cull et al. (1996) for schedules comprised of test trials only. By contrast, studies using distributed study trials or test trials followed by immediate feedback typically have not shown an advantage for an expanding schedule (Balota, Duchek, Sergeant-Marshall, & Roediger, 2006; Cull et al. 1996, Experiment 5; Cull, 2000).

Carpenter and DeLosh (2005) have suggested that an expanding retrieval schedule is optimal only when learners use a strategy that is focused on preventing already learned information from being forgotten. Test trials without feedback are assumed to facilitate a preventive maintenance strategy; therefore expanding schedules lead to superior performance when feedback is not provided. However, providing immediate feedback after each test trial is assumed to cause participants to shift to a strategy focused on correcting errors in learning, at the expense of maintaining or strengthening already learned information—thereby negating the effectiveness of the expanding schedule. To the extent that delaying feedback rather than providing immediate feedback encourages the use of a preventive maintenance strategy, our results are consistent with Carpenter and DeLosh's explanation for the advantage of an expanding retrieval schedule.

General Discussion

In a pair of experiments designed specifically to test theoretical explanations of feedback timing effects, we examined the effects of variations in feedback timing, post-feedback presentation, and inter-trial lag on the learning and retention of semantically rich materials. In Experiment 1, we observed a delay-retention effect in that providing delayed feedback led to better performance after a one-week retention interval than did providing immediate feedback. There was no evidence to support the hypothesis that the delay-retention effect is driven by a reduction in proactive interference from initially incorrect responses in the delayed feedback conditions, leading to reduced error perseveration and increased error correction probabilities. Instead, the effect was primarily driven by an increased probability of remembering initially correct responses, as predicted by a theoretical account that posits that feedback timing effects are due to the spacing of opportunities to process information. In Experiment 2, manipulating the inter-trial lag between study, test, and feedback trials lead to results consistent with those from studies of distributed practice, thus providing additional evidence to support the spacing hypothesis as an explanation of feedback timing effects. In the following sections, we discuss the implications of these findings for theories of feedback function and feedback timing, how these findings may help clarify some of the confusion in the feedback timing literature, and the implications of these findings for educational practice.

Theoretical Implications

The Role of Feedback in Learning from Tests

The assumption that feedback helps to correct errors rather than strengthening correct responses (Guthrie, 1971) has been a guiding principle of feedback research for the last 40 years (Pashler et al., 2005). While we do not claim that feedback does not correct errors, our findings, together with recent studies by Butler and colleagues (Butler et al., 2007; Butler & Roediger, 2008), show that, contrary to the Guthrian assumption, feedback can strengthen initially correct responses—at least when it is delivered after a moderate delay. A complete theory of feedback function necessarily should be able to explain the effects of feedback on initially correct responses as well as on incorrect and incomplete or blank responses. Our results thus suggest that theories of feedback function based on the Guthrian assumption are incomplete.

Theories of Feedback Timing Effects

Of course, this implies that theories based on the Guthrian (Guthrie, 1971) assumption cannot provide a general explanation of feedback timing effects, either. In both of our experiments, we found that while providing feedback did help correct initial errors, there was no systematic effect of feedback timing on error correction. In particular, we found no evidence of any beneficial effects of feedback delay on error correction and perseveration, contrary to predictions of the interference perseveration hypothesis.

Of course, it could be argued that we did not observe reliable differences in error correction effects because our participants exhibited relatively high learning rates on the initial test, especially in Experiment 1. If the initial learning rates were lower, and error rates higher, then error correction and interference effects might have had a larger impact on retention scores and spacing effects might have had a smaller impact. As an extreme example, in Kulhavy and

Anderson's (1972) study, participants had no exposure to the to-be-learned material prior to the initial test, and their mean scores on the initial test ranged from 26-30%, barely above chance on a 4AFC test. In situations such as this, the major function of feedback may well be to correct initial errors, but such error correction would be functionally equivalent to initial learning. In normal educational settings, participants are rarely tested on material to which they have not been previously exposed.

In contrast to the Guthrian (Guthrie, 1971) perspective, the spacing hypothesis (Butler et al., 2007; Smith et al., 2007) views test and feedback trials as additional processing opportunities that have the potential to strengthen responses through the same cognitive mechanisms that give rise to distributed practice effects. Thus, delaying feedback has two effects. First, it provides spaced retrieval practice trials for the items that learners get correct on the initial test, strengthening those items in memory and increasing the probability that they will continue to be remembered throughout the retention interval. Second, it increases the effective inter-trial lag (i.e., the study-feedback lag) for those items the learner did not get correct on the initial test. If this increase in lag is an appreciable fraction of the retention interval (say 5-20%), then this increase in lag could make the feedback trial more effective, thus increasing the probability that the correct response will be learned and retained.

Relationship to Other Feedback Timing Studies

Although our results are consistent with recent findings by Butler and Roediger (2008), they are only partially consistent with most previous feedback timing studies. As we mentioned in the introduction, we tried in our experimental designs to address a number of methodological issues that we think have led to disparate results in the literature. By addressing these issues, though, we also introduced differences in operational definitions, testing methods, and conditional analysis techniques that are likely to be responsible for the inconsistencies between our results and previous studies. We discuss each of these differences below.

Operational Definitions

The fact that standard operational definitions for feedback timing studies have not been established within the research community is perhaps the largest potential cause of inconsistencies between studies and the largest source of confusion in the feedback timing literature. For example, whereas some feedback timing studies have operationally defined immediate feedback as providing the correct answers after all items have been tested (e.g., Kulhavy & Anderson, 1972; Phye & Andre, 1989; Surber & Anderson, 1975; Swindell & Walls, 1993), others have defined this same timing manipulation as delayed feedback (e.g., Angell, 1949; Beeson, 1973; Brosvic et al., 2005; Rankin & Trepper, 1978). Naturally, this makes it difficult to compare the results from these studies.

In Experiment 1, we chose operational definitions of immediate and delayed feedback that closely matched theoretical constructs across three different feedback timing theories, and our results are consistent with previous studies that have used similar operational definitions. For example, Rankin and Trepper (1978) found that, relative to providing feedback immediately after each item, delaying feedback on a multiple choice test by about 10 minutes improved scores on a cued recall test taken after a 24 hr retention interval. Similarly, Sturges (1969, 1972, 1978) consistently found that, relative to providing feedback after each test question, delaying feedback by 20 mins or 24 hrs improved scores on 7-day retention tests in both laboratory and classroom studies using multiple choice tests.

In Experiment 2, we expanded our operational definitions beyond the binary "immediate vs. delayed" definitions and used operational definitions based on the amount of time (lag) between study, test, and feedback trials. This mirrors the recent transition in the study of distributed practice effects away from the binary "massed vs. spaced" definitions (Cepeda et al., 2006). Although we know of no previous feedback timing studies that have examined lag effects directly, a few have used different levels of delay that could be re-defined as lag manipulations for comparison to our results. For example, in a large study with junior high school students, More (1969) used four levels of test-feedback delay (immediately after each question, 2.5 hrs, 24 hrs, and 96 hrs) prior to giving the students a multiple-choice test after a 3-day retention interval. As the test-feedback lag increased from 0 to 24 hours, the scores on a retention test increased, but when the lag was increased to 96 hrs, the retention test scores began to decline. This pattern of results is similar to the non-monotonic lag effects that we observed in Experiment 2 and that are typically observed in distributed practice studies (Cepeda et al., 2006, 2009).

Conditional Analysis Techniques

By performing a complete conditional analysis that distinguished incorrect responses—errors—from blank responses and that further distinguished between new errors and perseverated errors, we were able to test the predictions of feedback timing theories more specifically than have previous feedback timing studies. Across both experiments, we found no evidence to support the two central predictions of the interference perseveration hypothesis—that delaying feedback reduces error perseveration rates and increases error correction probabilities. Although we used cued recall tests, several studies using multiple choice tests have also failed to find any differences in error perseveration as a function of immediate versus delayed feedback (Brosvic et al., 2005; Clariana et al., 2000; Peeck, 1979; Phye & Andre, 1989).

This raises a question: Why have other studies—in particular, the seminal studies by Anderson and colleagues (e.g., Kulhavy & Anderson, 1972; Surber & Anderson, 1975) that formed the basis for the interference perseveration hypothesis—found a reduction in error perseveration with a delay of feedback? One possible reason is that these studies have not adequately tracked item-specific errors. That is, they did not differentiate items in which the original error perseverated to the retention test (as would be expected based on the interference perseveration hypothesis) from items in which a new error was made on the retention test. Additionally, because these studies used multiple-choice tests in which participants had to respond to every question, even if it meant guessing, they were not able to distinguish between errors that were due to a lack of knowledge and errors that were due to incorrect knowledge.

Testing format

Although almost every feedback timing experiment reported in the literature has used a multiple-choice testing format (usually four alternative forced choice, 4AFC), we chose to use cued recall tests so that we could test the generalizability of feedback timing effects and theories, as well as avoid a number of potential problems associated with multiple-choice tests. The fact that we observed the delay-retention effect that is commonly found with multiple-choice tests shows that the effect can generalize to cued recall. However, there are two major differences between cued recall and multiple-choice tests that might influence the measurement and interpretation of feedback timing effects.

First, studies by Roediger and colleagues suggest that the presence of distractors could introduce a large number of potential confounds that could mask, emulate, or interact with feedback effects (Butler & Roediger, 2008; Kang, McDermott, & Roediger, 2007; Roediger & Marsh, 2005). For example, Butler and Roediger point out that the mere exposure to incorrect information can interfere with memory for related correct information. Thus, the distractors on a multiple-choice test might interfere with encoding of the correct answer, encoding of feedback, and even encoding of other distractors. Additionally, when Roediger and Marsh measured the effect of multiple-choice tests on learning, they found that performance on a final cued recall test was linearly related to the number of alternatives on an initial multiple-choice test, with correct recall decreasing and incorrect recall of multiple-choice distractors increasing as the number of alternatives increased. These results suggest that multiple-choice tests may be more prone to error perseveration effects than are cued recall tests.

Second, because multiple-choice tests are by definition forced-choice tests, participants have to guess whenever they do not know the answer to a test question. This guessing potentially contaminates conditional response probability analyses, especially for error perseveration probabilities, in at least two ways. For one thing, guessing adds noise to the probability of both initially correct and initially incorrect responses upon which conditional analyses are based, but because the probability of guessing correctly is smaller than the probability of guessing incorrectly, the noise is asymmetrical (Frary, 1980). Guessing also forces participants to commit to a response, which can then act to anchor or frame the question such that they are more likely to repeat the guess on a later test than they would be by chance alone (Marsh, Roediger, Bjork, & Bjork, 2007).

Implications for Education

In education, tests are often regarded as assessment tools rather than as learning tools (Dempster, 1996), despite the fact that the simple act of testing has been shown to have dramatic impacts on learning in classroom environments (Roediger & Karpicke, 2006) and that testing effects can be amplified even further when the learner is given feedback that includes the correct answer (Guthrie, 1971; Kulhavy, 1977). Of course, considering tests with feedback as learning events raises a question as to when the feedback should be provided in order to maximize its effectiveness. The results from our study suggest that the optimal timing for delivering feedback may be a complex function of the rate of initial learning, the study-test and test-feedback lags, and the desired length of retention. This function likely forms a “temporal ridgeline,” much as the optimal timing for repeated study trials does (Cepeda et al., 2008).

Despite this complexity, our results, together with others, suggest some guidelines for designing optimal feedback schedules. First, it is important to estimate the expected performance on the initial test. If a large proportion of errors are expected, then learners would appear to benefit from feedback delivered as soon after the test questions as is feasible in order to allow the learners to correct their mistakes (Guthrie, 1971). However, if the proportion of expected errors is relatively low, then delayed feedback would appear to be more beneficial because it leads to the strengthening of correct responses; in the context of classroom instruction, this could be at the next class meeting. In addition, the results from our manipulations of post-feedback processing suggest that whether feedback is provided immediately or at a delay, learners would benefit from restudying and retesting shortly after receiving feedback on their initial test performance (but see Fritz, Morris, Bjork, Gelman, & Wickens, 2000, for a discussion of when additional study and test trials are not helpful).

Conclusion

For over three decades, the widely held assumption that feedback following a test functions solely as an error correction mechanism went largely unchallenged (Mory, 2004). This study adds to a growing body of evidence that the strengthening of correct responses in memory is an equally important function of feedback. In the two experiments reported here, we consistently observed feedback timing effects that match the effects of distributed practice: Delaying feedback improved performance on a one-week retention test; the locus of this effect was an increase in correct response perseveration; varying the lag between study, test, and feedback trials gave rise to a non-monotonic lag effect curve; and expanding lags across those trials yielded retention performance superior to uniform or contracting lags. In summary, our results are consistent with the hypothesis that feedback timing effects are a particular class of spacing effects; by contrast, our results are inconsistent with theories assuming that the only function of feedback is to correct initial errors.

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.

Ausubel, D., Stager, M., & Gaite, A. (1969). Proactive effects in meaningful verbal learning and retention. *Journal of Educational Psychology*, *60*, 59-64.

Balota, D., Duchek, J., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging, 4*, 3-9.

Balota, D., Duchek, J., Sergent-Marshall, S., & Roediger, H. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging, 21,* 19-31.

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

Beasley, W. H., & Rodgers, J. L. (2009). Resampling methods. In R. E. Millsap & A. Maydeau-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology*. Thousand Oaks, CA: Sage.

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

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

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

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

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

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

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.

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.

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.

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

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236-246.

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19,* 1095-1102.

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

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

Cull, W., Shaughnessy, J., & Zechmeister, E. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied, 2*, 365-378.

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

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory.* (pp. 317-344). San Diego, CA: Academic Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1−26.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement, 4*, 79-90.

Fritz, C. O., Morris, P. E., Bjork, R. A., Gelman, R., & Wickens, T. D. (2000) When further learning fails: Stability and change following repeated presentation of text. *British Journal of Psychology, 91*, 493-511.

Glenberg, A. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1-16.

Guthrie, J. T. (1971). Feedback and sentence learning. *Journal of Verbal Learning and Verbal Behavior*, *10*, 12-17.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed). *Theories in cognitive psychology: The Loyola Symposium* (pp. 77-99). Potomac, MD: Erlbaum.

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

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

Kang, S., McDermott, K., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528-558.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook, Fourth edition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*, 211-232.

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

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

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

Manly, B. F. J. (2007). *Randomization*, *Bootstrap and Monte Carlo Methods in Biology*, *3rd ed.* Boca Raton, FL: Chapman & Hall/CRC.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*, 194-199.

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

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

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests, 2nd ed*. Mahwah, NJ: Erlbaum Associates.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005) When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 3-8.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*, 187-193.

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

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

Peterson, L., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology, 66*, 206-209.

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

Postman, L., Stark, K., & Burns, S. (1974). Sources of proactive inhibition on unpaced tests of retention. *American Journal of Psychology, 87*, 33-56.

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

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

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

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.

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

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

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

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

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Skinner, B.F. (1953). *Science and Human Behavior*. New York, NY: Macmillan.

Smith, T. A. (2007). *Spacing and the delay-retention effect: An alternative explanation of the effects of feedback timing on semantic learning*. Unpublished masters thesis, University of Texas at Arlington.

Smith, T.A. & Kimball, D.R., & Mann, M. (2007, November). *Feedback timing in semantic learning: A critical re-examination of the delay-retention effect*. Poster session presented at the 48th annual meeting of the Psychonomic Society, Long Beach, CA.

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

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

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

Sturges, P. T., & Crawford, J. J. (1964). The relative effectiveness of immediate and delayed reinforcement on learning academic material. (Research Report No. 05-02) Olympia, WA: Office of the State Superintendent of Public Instruction.

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

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

Appendix

Description of the Bootstrap Techniques Used in the Experiments

Bootstrapping is a well-established statistical technique for making inferential decisions using an empirical sampling distribution instead of a theoretical sampling distribution (Efron, 1979; see Beasley & Rodgers, 2009, or Manly, 2007, for recent treatments that discuss applications in psychology and the life sciences). The basic idea behind bootstrapping is that the distribution of population parameters can be estimated by repeatedly resampling the original sample, and that the resulting *bootstrap distribution* can then be used as the basis for calculating confidence intervals and *p* values, similar to the way in which these are calculated from theoretical sampling distributions in traditional inferential approaches. We chose to use the basic nonparameteric bootstrap method as described by Beasley and Rodgers to estimate the distribution of the relevant statistic for each comparison of interest under the assumption that the null hypothesis was true, as described below. In Experiment 1, the relevant statistics were the values from pair-wise contrasts between the immediate and delayed feedback conditions for error correction, error perseveration, and correct response perseveration probabilities. In Experiment 2, the relevant statistics were the values from quadratic contrasts to test for curvilinearity across three levels of lag for the probabilities of correct response perseveration and error correction.

To create the null hypothesis distributions for the relevant statistics, we first prepared a sampling frame containing the conditionalized item-level response data from each participant, in which the sampling units were the participants. On each iteration of the bootstrap algorithm, we created a bootstrap sample by drawing *N* samples with replacement from the sampling frame, where *N* was the number of participants in the experiment, and randomly shuffling the observed data from each sampled participant between the experimental conditions for the sample. The shuffling of data between the experimental conditions simulates the null hypothesis by removing whatever systematic differences might exist due to the experimental manipulations. After shuffling the data for each participant, we aggregated the item-level data for each experimental condition across the set of sample participants, calculated the mean of each dependent variable for each experimental condition, and calculated the contrast value for each contrast of interest. We repeated this resampling process for a total of 24,999 iterations for each experiment.

After creating the bootstrap distributions, we calculated two-tailed confidence intervals and *p* values for the observed contrast values, following the procedure recommend by Beasley and Rodgers (2009). Confidence intervals were calculated using the percentile method: The values in the bootstrap distribution were ordered, and the scores falling at the $\alpha/2$ and 1- $\alpha/2$ quantiles were used as the lower and upper limits, respectively. The *p* values were calculated using the following equation:

$$p = \frac{1 + \#\{|c^*| > |c_{obs}|\}}{B + 1},$$

where #{} is a counting function, c^* are the contrast values in the bootstrap distribution, c_{obs} is the observed contrast value, and *B* is the number of values in the bootstrap sample (24,999). As illustrations of the null-hypothesis distributions generated by these procedures, see Figure 2.

Author Note

A portion of the research reported here formed a portion of the first author's Master's thesis at The University of Texas at Arlington and was supported in part by Research Enhancement Program Grant # 14-7487-63 from UTA to the second author. Correspondence should be directed to D. R. Kimball, University of Oklahoma, Department of Psychology, 455 W. Lindsey St., Room, Norman, OK 73019 (e-mail: dkimball@psychology.ou.edu).

Footnotes

[1] Confidence rating data were collected for exploratory purposes. Because we opted to control the timing of the trials, we were not able to collect confidence judgments on every test trial. On average, participants gave confidence judgments on 77.4% of retention test trials, but when they did give a confidence rating, it tended to be high (*M* = 5.73, median = 7). A more detailed analysis of confidence rating data was reported in Smith (2007), and the data are available from the authors upon request.

[2] Previous studies that have examined conditional response probabilities have typically used between-subject designs with a large number of observations per participant, thus ensuring that every subject has at least one observation in each cell in the conditional response tree so that traditional inferential hypothesis tests can be conducted (e.g., Kulhavy & Anderson, 1972; Pashler et al., 2005). This approach has the principal disadvantage of requiring a large number of participants in order to obtain the power to detect even a moderate effect (Murphy & Moyers, 2004). For example, Pashler et al. used approximately 50 participants in each condition. A between-subjects version of our Experiment 1 would have required upwards of 400 participants. A secondary disadvantage of this approach that is rarely considered is that because the distributions of responses within the different cells are likely to have dramatically different moments (see Figure 3), using inferential tests such as the ANOVA that are susceptible to violations of heterogeneity of variance can lead to inaccurate conclusions unless appropriate adjustments are made (Keppel & Wickens, 2004, pp. 147-58).

[3] The authors would like to thank Will Beasley for assistance with the bootstrap analyses.

Table 1

Mean Proportion Correct Recall on the Retention Test for the Timing × Lag Conditions in Experiment 1.

	Feedback timing		
Study-test lag	Immediate	Delayed	*M*
Short	.54 (.022)	.65 (.022)	.60 (.016)
Long	.60 (.020)	.61 (.020)	.60 (.014)
M	.57 (.015)	.63 (.015)	

Note: Numbers in parenthesis are standard errors of the mean.

Table 2

Mean Proportion Correct Recall on the Retention Test for the Timing × Post-feedback Processing Conditions in Experiment 1

Post-feedback processing	Feedback timing		
	Immediate	Delayed	*M*
None	.54 (.022)	.65 (.022)	.60 (.016)
Restudy	.65 (.022)	.65 (.019)	.65 (.015)
Retest	.67 (.021)	.69 (.021)	.68 (.015)
M	.62 (.013)	.67 (.012)	

Note: Numbers in parenthesis are standard errors of the mean.

Table 3

Conditional Response Probabilities for Experiment 1

Initial test response	$C_{initial}$			$I_{initial}$				$B_{initial}$		
Retention test response	$\mathbf{C_{ret}}$	I_{ret}	B_{ret}	$\mathbf{C_{ret}}$	$\mathbf{I_{old}}$	I_{new}	B_{ret}	C_{ret}	I_{ret}	B_{ret}
Short study-test lag										
No re-processing										
Immediate feedback	**.68**	.14	.18	**.23**	**.17**	.26	.33	.11	.37	.52
Delayed feedback	**.83**	.08	.09	**.26**	**.15**	.19	.40	.16	.26	.58
Restudy										
Immediate feedback	**.80**	.08	.11	**.31**	**.12**	.27	.30	.22	.27	.51
Delayed feedback	**.83**	.09	.07	**.30**	**.11**	.19	.40	.23	.23	.54
Retest										
Immediate feedback	**.82**	.10	.08	**.44**	**.06**	.22	.28	.25	.23	.52
Delayed feedback	**.87**	.06	.07	**.30**	**.16**	.20	.35	.33	.28	.39
Long study-test lag										
No re-processing										
Immediate feedback	**.77**	.10	.12	**.29**	**.11**	.29	.31	.19	.22	.59
Delayed feedback	**.81**	.10	.09	**.23**	**.09**	.27	.41	.18	.23	.59

Note: C = correct response, I = incorrect response, I_{new} = new incorrect response, I_{old} = old incorrect response (error perseveration), B = blank (no response). Retention test response probabilities (C_{ret}, I_{ret}, I_{old}, I_{new}, and B_{ret}) are conditionalized on initial test responses. The columns for correct response perseveration ($C_{ret} | C_{initial}$), error correction ($C_{ret} | I_{initial}$), and error perseveration ($I_{old} | I_{initial}$) are in bold.

Table 4

Stimulus Presentation Schedule for the Learning Phase and Mean Proportion of Correct Responses on the Initial Test and the Retention Test in Experiment 2

Con-dition	Study-feedback lag (in days)	Practice schedule	Stimulus presentation schedule			Initial test (during learning)		Retention test (Day 8)	
			Day 1	Day 2	Day 3	*M*	*SE*	*M*	*SE*
1	0	Uniform	S-T-F			.89	.017	.71	.023
2	0	Uniform		S-T-F		.82	.017	.70	.025
3	0	Uniform			S-T-F	.85	.017	.72	.020
4	1	Expanding	S-T	F		.85	.016	.77	.019
5	1	Expanding		S-T	F	.85	.019	.78	.020
6	1	Contracting	S	T-F		.63	.022	.76	.021
7	1	Contracting		S	T-F	.62	.021	.73	.021
8	2	Uniform	S	T	F	.61	.023	.74	.024
9	2	Expanding	S-T		F	.86	.016	.79	.019
10	2	Contracting	S		T-F	.49	.020	.70	.025

Note: The stimulus presentation schedule for each of the 10 conditions is schematically represented in the center, where S is a study trial, T is a test trial, F is a feedback trial, and a single dash is a delay of approximately 8 minutes. The retention test for all conditions occurred on Day 8.

Figure Captions

Figure 1. Experimental design and stimulus presentation schedule for the learning session in Experiment 1. The 4 conditions inside the dashed square formed the *timing* × *lag* factorial manipulation, and the 6 conditions inside the solid square formed the *timing* × *post-feedback processing* factorial manipulation. The stimulus presentation schedule for each of the 8 conditions is schematically represented on the right, where S is a study trial, T is a test trial, F is a feedback trial, a single dash is a delay of 8-10 minutes, and a double dash is a delay of 16-20 minutes.

Figure 2. Bootstrap distributions for the difference between delayed feedback and immediate feedback for error correction, error perseveration, and correct response perseveration probabilities in Experiment 1. The spikes are the observed values. Portions of the distribution outside the two-tailed 95% confidence limits are shaded.

Figure 3. Idealized lag effect curves for short and long retention intervals.

Figure 4. Mean proportion of correct responses on the initial test and the retention test in Experiment 2 as a function of total lag across the study, test, and feedback trials.

Figure 5. Mean probabilities of a correct response on the retention test in Experiment 2 conditionalized on initial test performance, as a function of total lag across the study, test, and feedback trials.

Figure 1

Post-feedback presentation	Study-Test Lag	Feedback timing	
		Immediate	Delayed
	Long	1. S--TF	2. S--T-F
None	Short	3. S-TF	4. S-T-F
Restudy		5. S-TF-S	6. S-T-F-S
Retest		7. S-TF-T	8. S-T-F-T

Figure 2

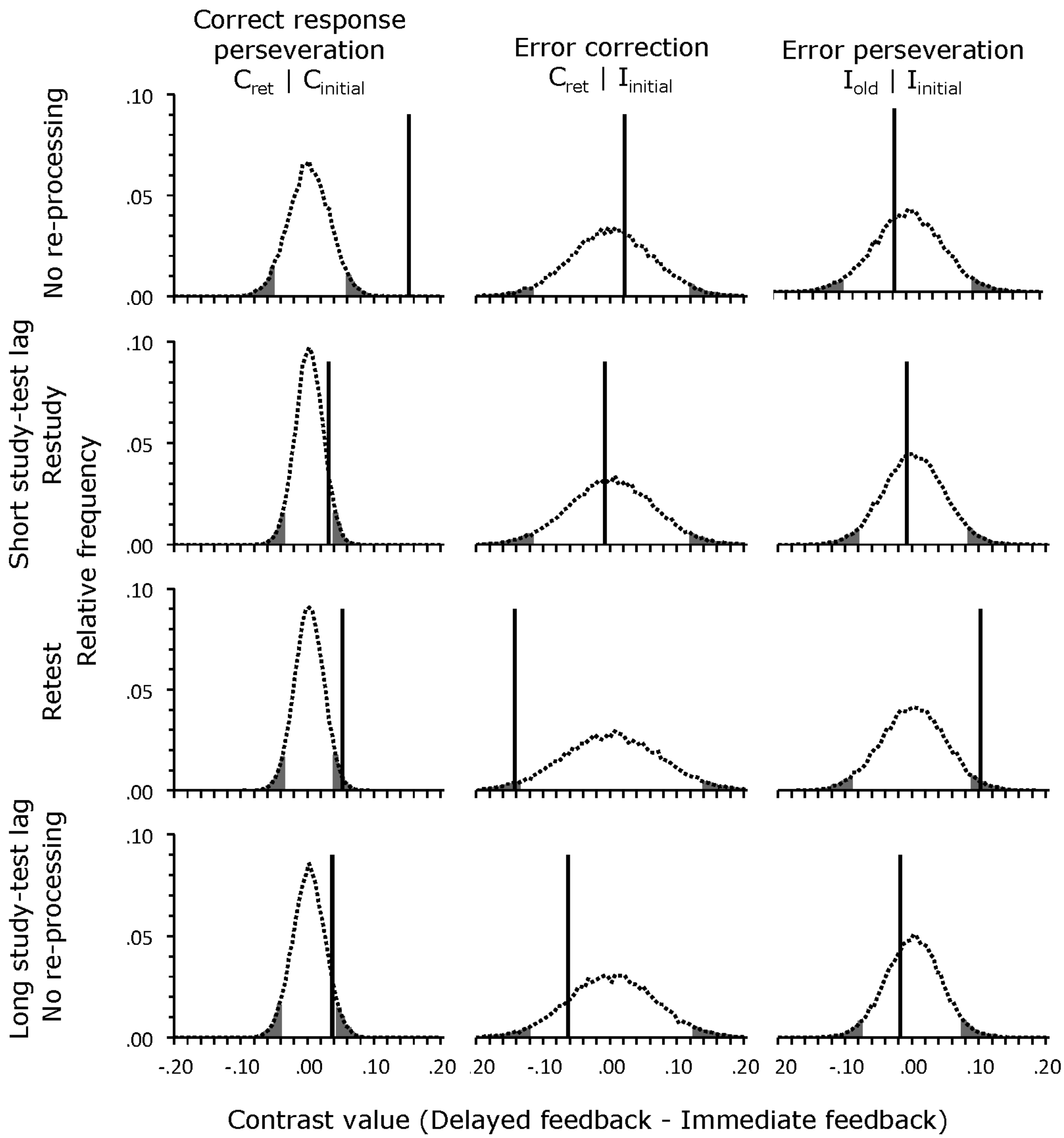

Correct response perseveration
C_{ret} | $C_{initial}$
Error correction
C_{ret} | $I_{initial}$
Error perseveration
I_{old} | $I_{initial}$
No re-processing
Short study-test lag
Restudy
Relative frequency
Retest
Long study-test lag
No re-processing
.10
.05
.00
-.20
-.10
.00
.10
.20
Contrast value (Delayed feedback - Immediate feedback)

Figure 3

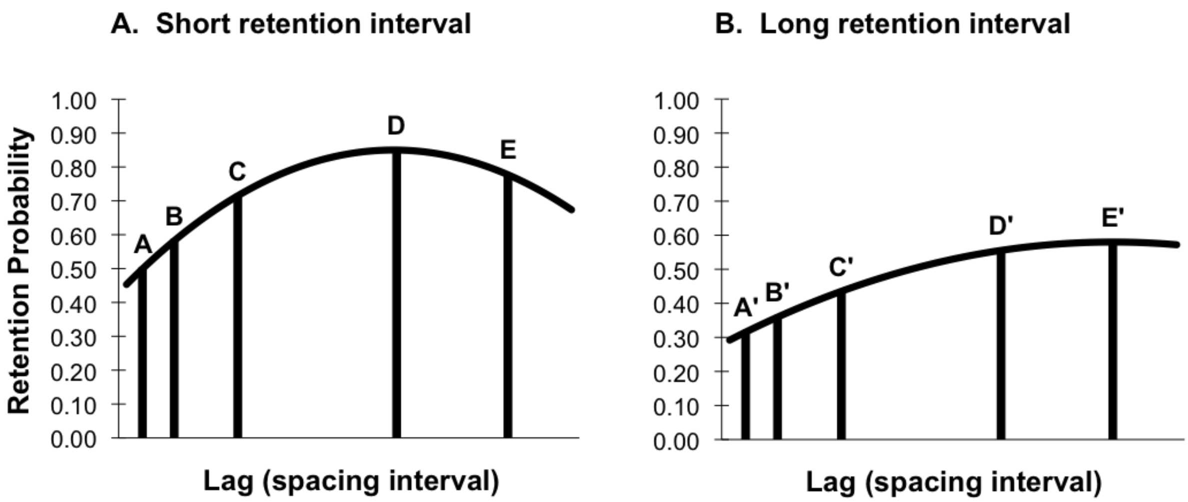

A. Short retention interval
B. Long retention interval
Retention Probability
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
A
B
C
D
E
A'
B'
C'
D'
E'
Lag (spacing interval)
Lag (spacing interval)

Figure 4

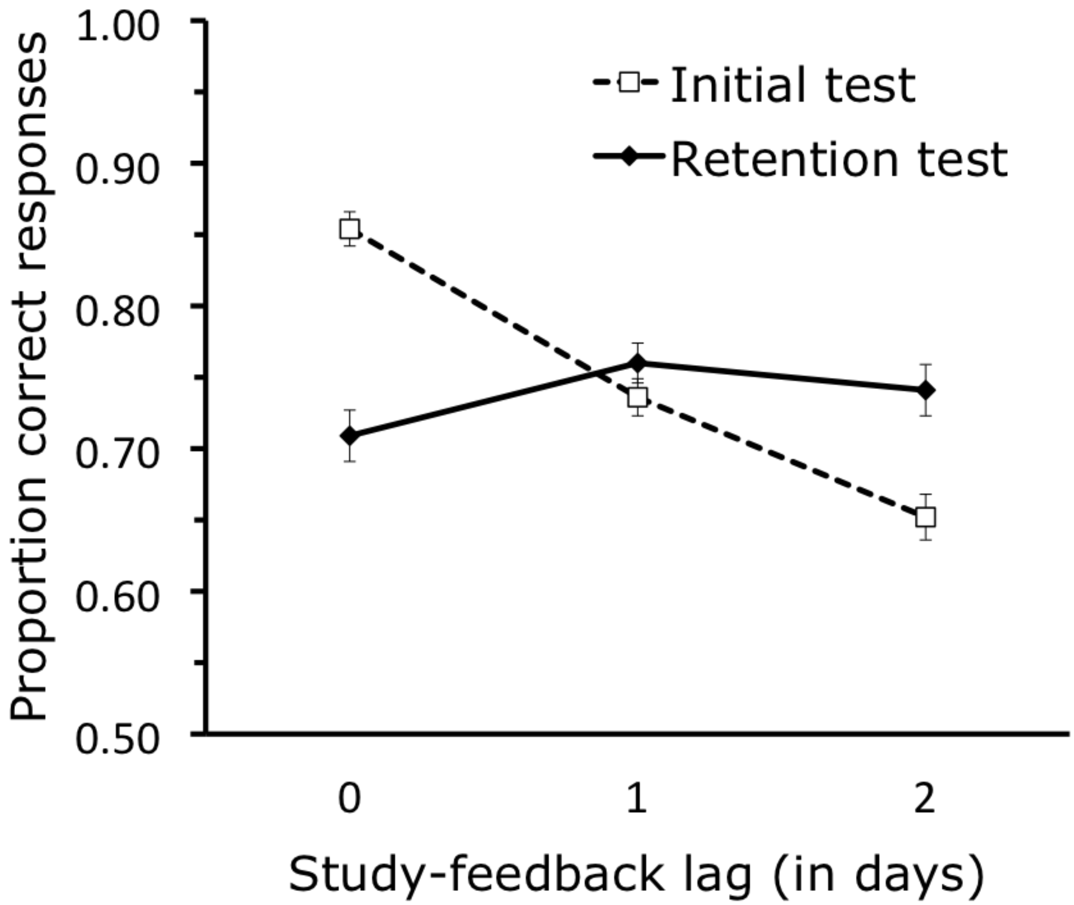
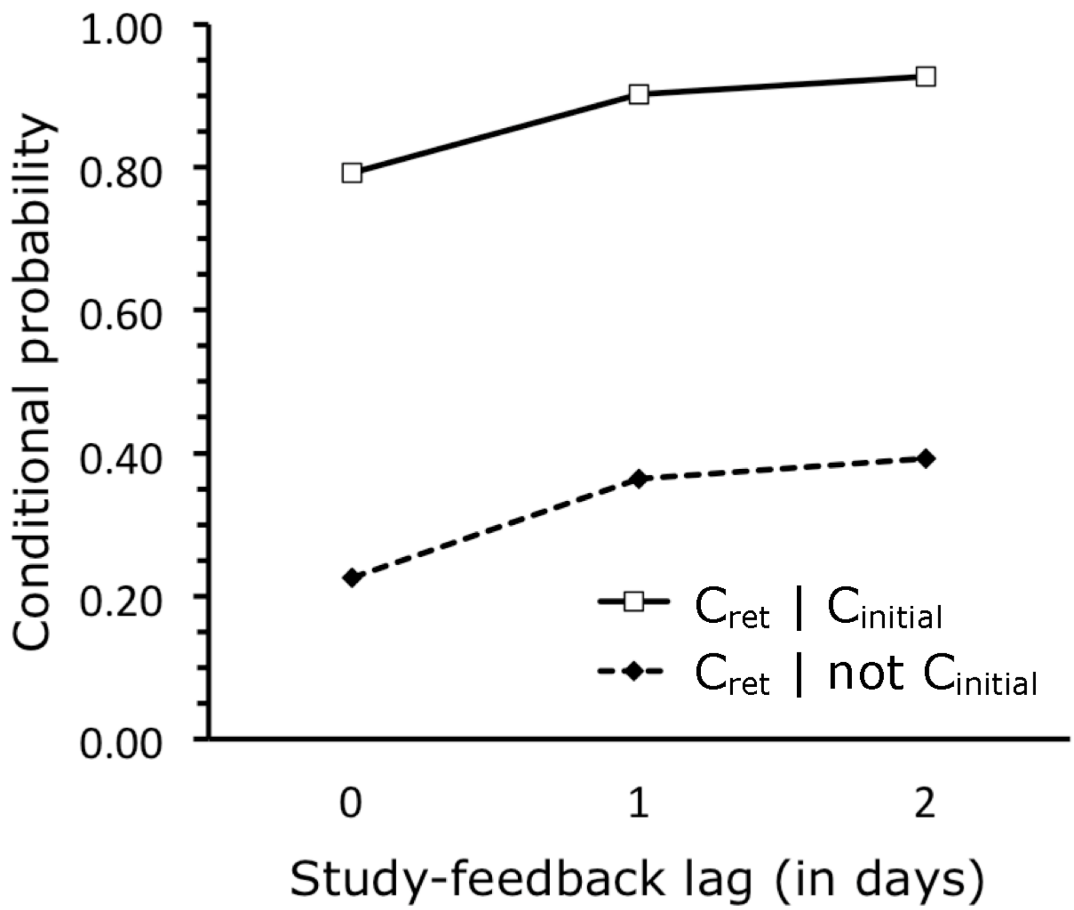

Proportion correct responses
1.00
0.90
0.80
0.70
0.60
0.50
Initial test
Retention test
0
1
2
Study-feedback lag (in days)

Figure 5


1.00
0.80
0.60
0.40
0.20
0.00
Conditional probability
0
1
2
Study-feedback lag (in days)
Cret | Cinitial
Cret | not Cinitial