Memory & Cognition
2009, 37 (8), 1077-1087
doi:10.3758/MC.37.8.1077

Delayed versus immediate feedback in children's and adults' vocabulary learning

Janet Metcalfe
Columbia University, New York, New York

Nate Kornell
Williams College, Williamstown, Massachusetts

and

Bridgid Finn
Columbia University, New York, New York

We investigated whether the superior memory performance sometimes seen with delayed rather than immediate feedback was attributable to the shorter retention interval (or *lag to test*) from the last presentation of the correct information in the delayed condition. Whether *lag to test* was controlled or not, delayed feedback produced better final test performance than did immediate feedback, which in turn produced better performance than did no feedback at all, when we tested Grade 6 children learning school-relevant vocabulary. With college students learning GRE-level words, however, delayed feedback produced better performance than did immediate feedback (and both were better than no feedback) when lag to test was uncontrolled, but there was no difference between the delayed and immediate feedback conditions when the lag to test was controlled.

The beneficial effects of corrective feedback on learning are now beyond dispute (R. C. Anderson, Kulhavy, & Andre, 1971; Butler, Karpicke, & Roediger, 2007, 2008; Butler & Roediger, 2008; Lhyle & Kulhavy, 1987; Metcalfe & Kornell, 2007; Metcalfe, Kornell, & Son, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005). When it is best to give that feedback, however, remains an unresolved question. It is this question that is addressed in the present article.

Some researchers, following behaviorist tenets that reinforcements need to be given quickly to be effective, have thought that it is essential to give feedback immediately. For example, Pressey (1950) stated,

> If test materials or simple testing devices could be developed such that, *as the student answered each question, that answer was immediately and automatically scored and recorded as right or wrong*, then clearly much trouble would be saved. Moreover, results would then *be available as soon as the test was finished*. . . . If he is weak on certain points, the test should locate them and aid in the remedying of these weaknesses. And this should be done *promptly*; an instructor who never answers a student's question until 48 hours after it is asked would be considered exasperatingly inefficient. The usual testing methods are grossly at fault in all these respects (p. 417).

Such immediate feedback has been implemented in state-of-the-art computer-based instructional technology called *cognitive tutors*, and is considered "the best tutorial interaction style" (J. R. Anderson, Corbett, Koedinger, & Pelletier, 1995, p. 167).

In most studies of human memory and learning of educationally relevant materials, reinforcement principles are not at issue. Even so, a compelling argument for why immediate feedback might result in superior performance can be made: If an error is allowed to stand uncorrected, it may be rehearsed, consolidated, and strengthened and may be more likely to recur than if it were immediately corrected. If feedback is given immediately, the correct answer, rather than an error, can then be rehearsed and consolidated.

A number of studies have reported better performance when feedback was given immediately. For example, Kulik and Kulik (1988) reported a meta-analysis of 53 studies that varied widely in the methodologies used. The conclusion to the meta-analysis was that although delayed feedback was often found to produce better results in laboratory studies, immediate feedback resulted in better performance in applied studies in actual classrooms, and Kulik and Kulik implied that it might be the classroom setting itself that was the key factor. Butler et al. (2007) protested that the breakdown of studies into those done in the classroom and those done in the laboratory is unsatisfying as an explanation and instead suggested that there might have been a difference in the learners' processing of the feedback in these studies. They noted

J. Metcalfe, jm348@columbia.edu

that, in many of the classroom situations, the learners may not have paid as careful attention to the feedback when it was given at a delay as when it was given immediately and that this difference in processing of the feedback may have accounted for the sometimes-seen superiority of the immediately given feedback. Laboratory studies, being better controlled, were less susceptible to this criticism.

It is, of course, important that studies investigating differences in the timing of feedback take special care to ensure that the feedback is processed and attended to by the participants to the same extent whether that feedback is given immediately or at a delay. In the experiments that follow, we had the participants type the feedback into a computer in both the delayed and immediate conditions, rather than simply passively viewing or hearing the feedback. The first experiment was conducted in the classroom, however. If the classroom setting itself was the reason for the superiority of immediate feedback, immediate feedback superiority should also be found in our first experiment. If differences in attention to the feedback were at the root of the sometimes-seen superiority of immediate feedback in the studies that Kulik and Kulik (1988) reviewed, we would not necessarily expect that immediate feedback would be better.

The meta-analysis of Kulik and Kulik (1988) also revealed that delayed feedback sometimes resulted in superior performance, and several other lines of research underline this possibility. For example, Guzman-Muñoz and Johnson (2007) showed that in learning geographical representations, delayed feedback—which entailed seeing an entire map, including the relations among to-be-learned places—resulted in a more laborious acquisition but better eventual retention than did immediate feedback on a test of the location of individual places. Guzman-Muñoz and Johnson's result, however, might have obtained not because of the delayed feedback per se, but rather because the configural information—which was helpful to performance in this task—was more salient in the delayed than in the immediate feedback case, as Guzman-Muñoz and Johnson suggested. Other researchers have proposed other reasons for why delayed feedback might enhance later memory. Bjork and Linn (2006) proposed the idea that processing difficulties at the time of encoding can enhance memory. The processing of delayed feedback may be more difficult than the processing of immediate feedback. Butler et al. (2007) pointed to differences in the spacing of the to-be-learned materials that obtain between immediate and delayed feedback conditions. The repetitions of the information with immediate feedback tend to be massed, whereas those with delayed feedback tend to be more dispersed or spaced. Insofar as spaced practice can benefit memory, as has often been shown (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a review), delayed feedback should benefit memory.

Kulhavy and Anderson (1972; see also Kulhavy, 1977) found that delayed feedback resulted in better eventual performance than did immediate feedback in 8 of the 11 experiments that they surveyed, as well as in their own reported study. Following Brackbill, Bravos, and Starr (1962), Kulhavy and Anderson called the observed superior performance with delayed feedback the delayed-retention effect (DRE). The explanation that Kulhavy and Anderson forwarded for the DRE was that learners purportedly forgot their incorrect answers during the delay interval when feedback was delayed but did not do so if feedback was immediate. Because of the increased forgetting of the wrong answer in the delayed condition, there was less postulated proactive interference from those incorrect answers at the time the feedback was given. With less proactive interference, the new correct answer could be more easily learned and better remembered.

Although Kulhavy and Anderson's (1972) rationale for their own results and those of the experiments that they surveyed lends credence to the idea that delaying feedback could have beneficial consequences, the design of the experiments that contributed to this conclusion invites a different explanation. The design is shown in Figure 1. As can be seen from the figure, the total time from the initial test to the final test was held constant in these experiments. Between these two test anchors, what varies is when the feedback was given. In the immediate feedback condition, the feedback was given virtually at the time of the initial test; in the delayed feedback condition, it was given somewhat closer to the time of the final test. But this means that the time between the last presentation of the correct answer (in the form of feedback) and the final test was shorter in the delayed feedback condition than in the immediate feedback condition: The delayed feedback condition enjoyed a shorter retention interval from the last presentation of the correct answer than did the immediate feedback condition. It is well established that the retention interval—or the *lag to test*—is an important determinant of memory performance (see, e.g., Murdock, 1974). This factor alone could have caused the DRE.

Although they controlled time from the initial test to the final test, but not time from feedback to the final test, in the

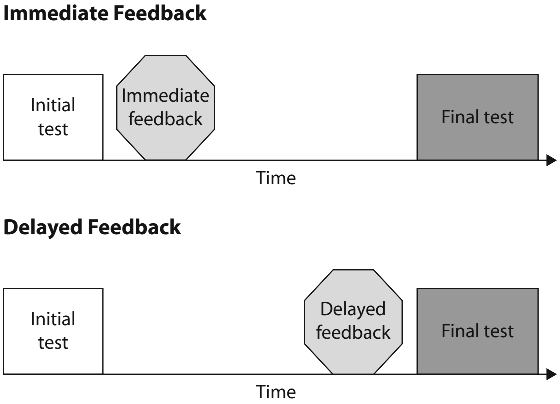

Figure 1. The design of experiments showing the delayed-retention effect. Although the time between the initial test and the final test was controlled, the time between the presentation of feedback in the immediate and delayed conditions and the final test was not equivalent.

experiments described by Kulhavy and Anderson (1972) even modern experimenters have not investigated this possible locus of the DRE. Differential lag to test may have contributed to some of the results. For example, in Experiment 1 reported by Butler et al. (2007), immediate feedback was given right after reading the to-be-remembered passages, whereas delayed feedback was given only 10 min later, and the final test was conducted at a 1-day interval (i.e., 1,400 min later). The ratio for the time from immediate feedback to the final test to the total from initial test to the final test (1,400 min/1,400 min = 1.00) and that for the time from delayed feedback to the final test to the initial test to the final test (1,390 min/1,400 min = .99) were almost identical. This experiment failed to demonstrate a significant DRE. In their second experiment, however, these lag-to-test ratios differed more substantially. As before, the ratio for immediate feedback was close to 1.00. The delayed feedback in this second experiment was given 1 day following the initial test, with the final test being 7 days following initial test, resulting in a ratio of .86—a much greater difference than in the first experiment—and a significant DRE was produced. Although there are exceptions (see Butler & Roediger, 2008), it seems plausible that much of the advantage of delayed feedback over immediate feedback that has been documented in past studies was attributable to a difference between conditions in lag to test. If the lag to test were constant, it is possible that no delayed feedback effect would occur, or even that immediate feedback would produce superior results.

EXPERIMENT 1

To investigate the timing of feedback, we needed a design in which the lag to test was the same in the immediate feedback and delayed feedback conditions. We also sought to ensure that the feedback itself was processed fully in both feedback conditions. Accordingly, we designed a four-session experiment in which the sessions were spaced a day or two apart. There were three sets of to-be-learned materials. After studying and being tested on a set of questions, the participants were given immediate feedback (which they then had to type into the computer themselves) on some of the questions on which they had made errors, delayed feedback on some (which, again, had to be typed in), and no feedback on some. The delayed feedback for the questions on which the students had made errors in the first session, however, occurred a few moments before the immediate feedback given to the errors that the students had made to questions on the second session, and the delayed feedback to the errors that the students had made during the second session was given during the third session, just before the immediate feedback to the errors made on the third session. Finally, on the fourth session, all items that had been incorrect were tested. By comparing performance on the questions that had been given feedback on Session 2 (i.e., items that were Session 1 questions being given delayed feedback, as well as items that were Session 2 questions being given immediate feedback), we could look at the effect of immediate versus delayed feedback while keeping lag to test constant. We could do the same thing with the questions that were given feedback on Session 3, some of which were items presented in Session 2 being given delayed feedback and some of which were items presented in Session 3 being given immediate feedback. We were also able to investigate the effect of immediate and delayed feedback when the time from the original event was held constant but there was a lag-to-test artifact, as in the previous experiments that showed the DRE.

Our hypothesis was that when the lag to test was allowed to vary, as in previous experiments, we would find better performance under delayed feedback than under immediate feedback conditions. The delayed feedback items would be the items that, on average, would have been presented for the last time closer to the test, and there would hence be a shorter retention interval. We predicted a diminution or even reversal of this effect when the lag to test was held constant. A reversal, with immediate feedback being superior, would be consistent with the view that immediate feedback is best in a classroom situation. It is also consistent with the idea that errors that were not immediately corrected might be strengthened or consolidated in the delay interval and might pose overwriting problems or interference with the correct answer. A diminution in the DRE, but one that still resulted in delayed feedback being better than immediate feedback, would be consistent with the idea that the delayed feedback condition would benefit from the greater spacing between the first presentation of the correct answer and the feedback and also with the possibility that the delay interval would result in forgetting of the errors and a decrease in proactive inhibition on the learning of the correct response. Thus, we predicted a diminution but did not know whether the DRE would reverse or not.

In Experiment 1, we focused on grade school children who were learning materials needed for their social studies class in a classroom setting. The reason for our focus on children in this experiment was that the importance of memory-enhancing variables may be more pronounced in this group of learners than it is in older and more sophisticated learners. We have found in previous studies (e.g., Metcalfe, 2006) that manipulations that enhance memory sometimes have a larger effect on grade school children than on college students. The latter can and often do compensate for shortcomings in the presentation of the materials and enact effective study strategies on their own. Grade school children are less likely to do so, leaving more of the onus on the teacher (or experimenter). Additionally, the main practical value of our study may be in grade school settings. Thus, our central interest was to investigate these effects with grade school children. In Experiment 2, however, we extended the research to college students.

There were three conditions: immediate feedback, delayed feedback, and no feedback. The experiment took place over the course of four sessions, with the final test occurring on the fourth. An overview of the design is shown in Figure 2.

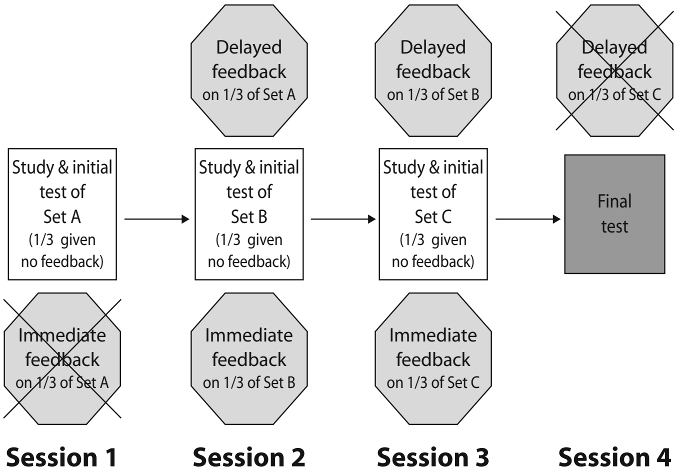
Delayed feedback on 1/3 of Set A
Delayed feedback on 1/3 of Set B
Delayed feedback on 1/3 of Set C
Study & initial test of Set A (1/3 given no feedback)
Study & initial test of Set B (1/3 given no feedback)
Study & initial test of Set C (1/3 given no feedback)
Final test
Immediate feedback on 1/3 of Set A
Immediate feedback on 1/3 of Set B
Immediate feedback on 1/3 of Set C
Session 1
Session 2
Session 3
Session 4

Figure 2. The design of the present experiments, in which the delay to test was controlled. The feedback sessions that have an "X" through them are the ones that are not included in the unconfounded analysis but that are included in the confounded analyses given in this article.

Method

Participants

The participants were 27 students enrolled in Grade 6 at The School at Columbia University in New York City. About half of the students were the children of university faculty and staff; the other half were from the local neighborhoods of Morningside Heights and Harlem.

Instructions and Interface

The experiment was run on Macintosh iBook computers using a computer program that presented learning as a fun and exciting game, which we called *Dragon Master*. The game aspect of the program was designed to motivate participants, without hindering their learning at a cognitive level. The story behind the game was that the participant was a dragon who started as an egg and grew larger and more powerful (i.e., advanced through the levels of the game) by answering questions correctly. The background on the screen appeared to be the inside of a cave. Correct answers were rewarded with coins that were displayed on the screen and that could be converted into pots of gold; when enough pots of gold were accumulated, the dragon advanced to a new level. The levels included hatching, developing wings, learning to fly, breathing fire, and so forth. An image of the dragon at its current stage of development was displayed on the screen. The instructions were presented on the screen, and at the same time, a recorded voice spoke the instructions aloud. The voice explained the story and said that he was Merlin, the dragon's wise guide. There was ominous background music during the instructions, as well as a picture of Merlin. The children were excited about earning the coins and pots of gold, as well as progressing to various dragon levels, and enjoyed the game-like quality of the experiment.

Procedure

There were four experimental sessions. The first session was on a Thursday; the second session was 1 day later, on Friday; the third session was 4 days later, on the following Tuesday; and the fourth session was 1 day later, on the immediately following Wednesday. Thus, the four sessions took up four consecutive periods of a regularly scheduled social sciences class during normal class time. The sessions were conducted in two classrooms (which were the home rooms for the participants). The time of day was constant across sessions. Learning and feedback took place during the first three sessions; the fourth session consisted of a delayed test.

Twenty-four new words were introduced during each of the three learning sessions (Sessions 1–3). During the learning sessions, there were four phases: delayed feedback from the previous session (which did not occur in Session 1), presentation of the new definitions, initial test on the new definitions, and immediate feedback on one third of the new definitions (randomly selected). Instructions explaining each phase were presented before the phase began.

Learning phase. The first phase of the experiment was the presentation phase. During the presentation phase, the 24 new words for the session were presented, one at a time. Each word was shown for 2 sec. At the same time, a recording of the word being spoken was played aloud over the participant's headphones. The definition was then presented for 6.5 sec, during which time a recording of the definition was played aloud and the word being defined remained onscreen. Then the recording of the defined to-be-remembered word was played again for another 2 sec.

Initial test. Initial learning of all 24 items was followed by an initial test. During the initial test phase, the participants were presented with each definition, one at a time. They were asked to type in the target word. If their responses were correct, they received a gold coin. If they had enough gold coins to convert to a pot of gold, this occurred onscreen, and if they had enough pots to advance a dragon level, this also occurred onscreen as soon as the critical correct response was typed in. The items that the participants answered correctly during the initial test were removed from the rest of the experiment; that is, they were not presented or tested again, nor were they included in the data analysis. By removing these items, we ensured that feedback would be manipulated only for items that the participants could not answer correctly during the initial test.

Feedback. After the entry of each item in the initial test, the items that had been incorrect were randomly divided into three feedback conditions: delayed feedback, immediate feedback, and no feedback. The items were assigned randomly in equal numbers to each condition. The initial test phase was followed directly by the immediate feedback phase. As in the initial study phase, each word was shown for 2 sec, accompanied by a recording of the word being spoken aloud, and then the definition was presented for 6.5 sec, accompanied by a spoken, computerized 2-sec sound bite of the answer, after which the word, which was already showing, was spoken aloud again for 2 sec. Then the word disappeared, and the participant was asked to type it in. If they typed it in correctly, they went on to the next item; if they did not type it in correctly, the word reappeared and was spoken aloud for another 2 sec, and then it disappeared and the participants were again asked to type it in. This process repeated until the correct answer was typed in. After all of the designated items had received feedback, the feedback cycle was repeated a second time. The mean number of attempts that were made by the children to the correct answer during this feedback phase was 1.49 on the first cycle of the feedback and 1.23 on the second cycle of the feedback. The very few errors until correct were largely spelling or typing errors. The requirement that the participants type in the correct answer was designed to ensure that the participants were actively paying attention and that they encoded the feedback correctly.

During Session 1, there was no delayed feedback phase, of course. Sessions 2–4, however, began with delayed feedback on the items studied during the previous session. The procedure used to give delayed feedback was the same as the procedure used to give immediate feedback, except, of course, that it occurred at the beginning of the next session.

Final test. The final test occurred during Session 4. Session 4 began with delayed feedback on the items assigned to the delayed feedback condition in Session 3, after which the final test began. The final test consisted of all of the items assigned to the delayed test condition. Each definition was presented, one by one, and the participants were asked to type in the corresponding word.

Materials

The materials were 72 vocabulary words, supplied by the Curriculum Director of The School at Columbia University, Marc Meyer, who designed the Grade 6 social science course and the corresponding materials for our study. The materials were very difficult (for Grade 6 children) vocabulary items related to a unit that the students were studying about Mexico (although the definitions were not related to the Spanish language). For example, two of the word–definition pairs were *inefficiency*—"performing tasks in a way that is not organized or failing to make the best use of something, especially time"—and *inscription*—"words or letters written, printed, or engraved on a surface."

Setting

The experiments took place over the course of four consecutive regularly scheduled social science classes in the students' classroom, during school hours. The school provided each student with a Macintosh iBook computer to use throughout the year. The software used to conduct the experiment was loaded onto the students' own computers before the experiment began. The students' teacher introduced the experimenters at the beginning of the first session but was not present during the experiment; there were 2 or 3 adult assistants in each of two rooms, including 1 or more of the authors and a research assistant, during the experiment. The 27 students were divided between two classrooms. Each student wore headphones, and distractions during learning and testing, although not nonexistent, were minimal.

Results

Exclusions

Data from 9 of the 27 participants were excluded, either because the participants missed multiple sessions because of absences from school or because they accessed the program on their computers at times other than the class times in which they were supposed to be doing the experiment, as had been designated by the experimenters. Some students were highly motivated by the Dragon Master game and played the program during recess and breaks. The experimenters could not control this, since they were at the school only when administering the experiment, but the children had access to their own computers at other times. Unfortunately, we were unable to use their data if they did this. (We were able to see that they had gone into the program between official sessions because all entries on the computer were time stamped.) Of the 18 participants whose data were included, 12 completed all four sessions as planned; the other 6 completed three sessions (two consecutive study sessions and a test session). Since lag to test could be controlled in the 6 participants who completed only three sessions, for the sessions that they completed, and since all analyses were within participants, their data were included. We computed the proportion correct for them on the basis of only the items that they had studied and on which they were given feedback. There were no significant differences between the participants who had completed only two study sessions and the participants who had completed all three study sessions.

In all of the experiments reported here, answers on the final test that were almost correct but contained misspellings were counted as correct. Final test accuracy was analyzed using a lenient scoring letter-match algorithm developed by Brady Butterfield, which corresponds to what independent scorers would usually consider to be correct but perhaps a spelling mistake. The results were the same, however, when we used a strict scoring algorithm requiring that the answers be letter perfect.

Because some items were given delayed feedback and others were given immediate feedback or no feedback, there were differences between the conditions with respect to the lag between an item's final study opportunity and the final test, which we simply collapsed over sessions. For example, items initially studied in Session 3 that were assigned to the immediate feedback or no-feedback condition were last studied during Session 3; items initially studied in Session 3 that were assigned to the delayed feedback condition were last studied in Session 4. Thus, we analyzed the data in two different ways: not controlling for lag to test and controlling for lag to test.

Uncontrolled for Lag to Test

In this first analysis, all items that had been incorrect and hence assigned to the immediate feedback condition (including those initially tested and given feedback on Session 1), had been assigned to the delayed feedback condition (including those initially tested on Session 3 and for which feedback was given on Session 4, just before the test), or had been assigned to the no-feedback condition were included in the analysis. The inclusion of all of the items meant that the average lag to test was greater in the immediate feedback (4.16 days, $SE = .24$) than in the delayed feedback condition (2.20 days, $SE = .16$). When the data were analyzed thus, without controlling for lag to test, as is shown in Figure 3A, there was a significant effect of feedback condition [$F(2,34) = 39.13, p < .0001, \eta_p^2 = .70$], as was expected. Post hoc Tukey tests revealed that performance in the delayed feedback condition was superior to performance in both the immediate and no-feedback conditions.

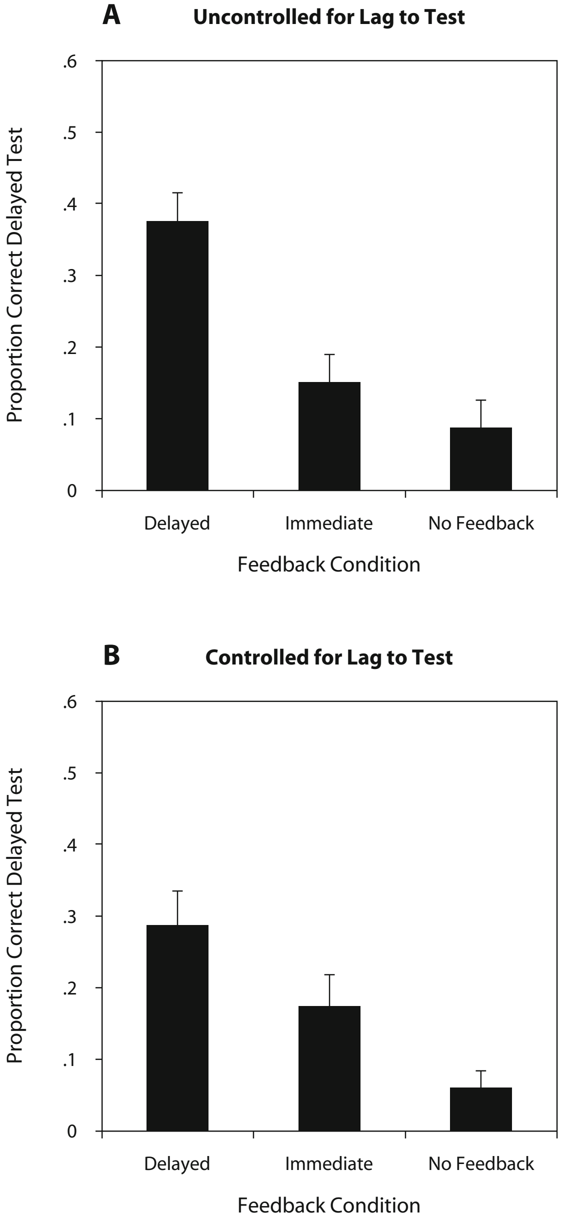

Figure 3. (A) Proportion correct as a function of feedback condition when lag to test was not controlled, with Grade 6 children. (B) Proportion correct as a function of feedback condition when lag to test was controlled, with Grade 6 children. Error bars represent standard errors.

Controlled for Lag to Test

To control for the differential time to test, we first excluded items studied in Session 3 that were assigned to the delayed feedback condition and so were last studied in Session 4. We also excluded items studied in Session 1 that were assigned to the immediate feedback or no-feedback conditions. The data were then analyzed in two ways. First, we conducted an ANOVA treating session (either 2 or 3) as a factor. *Session*, in this analysis, meant the session on which the feedback was given, which for immediate feedback was also the session on which the items had first been studied and tested, but for delayed feedback was the session that followed the session on which the item had originally been studied and tested. Unfortunately, there were many empty cells when the data were treated in this way, because there were not a sufficient number of incorrect responses to provide data in each of the three feedback conditions in each of the two sessions, and so the data from only 11 participants qualified for the analysis. The effect of session was not significant ($F < 1$), nor was the interaction between session and feedback significant ($F < 1$). The effect of feedback, however, was significant [$F(2,20) = 6.21$, $MS_e = .03$, $p < .008$, $\eta_p^2 = .38$]. The mean recall for delayed feedback was .29 ($SE = .05$); for immediate feedback, it was .17 ($SE = .05$); and for no feedback, it was .06 ($SE = .03$).

In the second analysis, because session had not been significant in the previous analysis, we collapsed over session. This allowed the inclusion of data from all of the participants. We computed the mean number of days between the last presentation of the correct answer, in the form of feedback, and the final test for both the immediate and delayed feedback conditions. The mean number of days from feedback to test was 2.44 ($SE = .26$) in the immediate feedback condition and 3.13 ($SE = .31$) in the delayed condition. These were not significantly different [$t(17) = 1.44$, $p > .05$]. Being reassured that collapsing did not introduce a difference in lag to test between the two critical conditions, we then conducted an ANOVA on the collapsed data. The effect of feedback (delayed feedback, immediate feedback, or no feedback) was significant [$F(2,34) = 12.37$, $p < .001$, $\eta_p^2 = .42$]. As can be seen in Figure 3B, when no feedback was given, performance was extremely poor. Immediate feedback improved performance substantially, but delayed feedback improved it even more. Post hoc Tukey tests revealed significant differences among all three conditions: Final test accuracy was highest in the delayed feedback condition, which was significantly higher than recall in the immediate feedback condition, which was significantly higher than recall in the no-feedback condition. Thus, in this experiment, we found a beneficial effect of delaying feedback, regardless of whether lag to test was controlled.

Discussion

Delayed feedback was superior to immediate feedback in this experiment. Although the effect was larger when lag to test was not controlled, it was still a large and significant effect even when we controlled for lag to test. Thus, the superiority of delayed feedback in this experiment was not due only to a differential lag to test, as might have been the case in previous experiments.

It might be objected that in our experiment, the participants almost certainly and immediately knew that their incorrect responses—including those assigned to the immediate condition, the delayed condition, and the no-feedback condition—were incorrect. Although they were not explicitly told this, they were given highly salient positive feedback—in the form of gold coins dropping and dragon levels changing—when they made a correct response. When no gold coins dropped, they could infer that they had given an incorrect response. Pashler et al. (2005) called this labeling of whether a response was correct or incorrect *partial* feedback. It is conceivable that had we given no feedback at all, rather than such partial feedback, until the immediate or delayed corrective feedback was given, our results might have been different. Although this is, of course, possible, we think that it is unlikely. Pashler et al. included treatment combinations in which no feedback was given and partial feedback was given. Partial feedback—the participants being told that answers were correct or incorrect—had no effect whatsoever. In their experiment, it was only when feedback in the form of the correct answer was given that there was a beneficial effect. Thus, although we cannot definitively rule out the possibility, we have no principled reason to suppose that partial feedback qualified our results. In Experiment 2, however, we were able to further investigate this possibility because one condition was just like those in Experiment 1 and had partial feedback, whereas the comparison condition did not use the Dragon Master game, and so did not include partial feedback.

Two explanations of the results of this first experiment seemed viable. The first was that given by Kulhavy and Anderson (1972) concerning errors. As was suggested by Kulhavy and Anderson as the reason for the DRE, the children may have been more likely to have forgotten their incorrect answers during the days that followed the initial test in the delayed condition than if no delay had been interposed. The incorrect responses—because of this increased forgetting—might have interfered less with the acquisition of the correct responses in the delayed condition than they would have had the feedback been given immediately, when those incorrect responses were still fresh in the mind.

The second explanation for the enhanced performance seen with delayed feedback was that given by Butler et al. (2007)—namely, that with delayed feedback, there was a greater spacing from the original presentation of the correct answer to the second correct presentation that was given at the time of feedback. In the immediate feedback condition, the repetition of the correct answer was similar to massed practice, whereas in the delayed feedback condition, it was more like spaced or distributed practice. Many studies have shown the beneficial effects of distributed practice (see Rohrer & Pashler, 2007, and Son, 2004).

EXPERIMENT 2

We conducted a second experiment, like the first one but with four differences. First, the participants were college students rather than Grade 6 children. Second, the materials were appropriate to college students—definition–word pairs that were at approximately the level of difficulty of the vocabulary tested in the GRE (or possibly more difficult)—rather than definitions used in a Grade 6 social studies curriculum. Third, whether the learning was conducted within the Dragon Master game shell, described above, was manipulated. For half of the participants, the exact same program as had been used with the children was used, except that the new materials were inserted into the program. For the other half of the participants, the Dragon Master game shell, including the music, the various game levels, the Merlin wise guide prologue, and the dropping of gold coins with correct responses, was simply eliminated and replaced by plain text instructions on the computer. The no-game participants underwent the same procedure and timing of study and feedback (including the entry of the responses, as well as the spoken words and definitions) as did the participants who had the game frame. We had been unable to test the effects of the Dragon Master game itself with the children, because they were so engaged by the game that it would have been perceived as punitive to test some of them without it. Including a no-game condition also allowed us to determine whether having the feedback at the time of responding had any effects, since in the no-game condition, no gold coins or other indications occurred at responding. We did not think this had an effect, but the second experiment allowed us to check. The fourth main difference was that the experiment was conducted in our laboratory rather than in a classroom setting.

Method

The participants were 20 Columbia University students who were paid for participating. The materials were 75 difficult vocabulary word–definition pairs, given in the Appendix. At the beginning of each session, 25 rather than 24 word–definition pairs (as had been the case in Experiment 1) were presented for study and subsequent test. The design was otherwise identical to that of Experiment 1, except that there was also a between-participants factor, Dragon Master (game or no game). The mean gap between Sessions 1 and 2 was 3.85 days; between Sessions 2 and 3, it was 3.85 days; and between Sessions 3 and 4, it was 2.35 days. The mean number of attempts that were made by the adults to produce the correct answer during the first cycle of feedback was 1.12 and was 1.04 on the second cycle of feedback; that is, most attempts were correct on the first try.

Results

In the first analysis, we included Dragon Master (game or no game) as a factor. Neither the effect of Dragon Master ($F < 1$) nor the interaction between Dragon Master and feedback [$F(2,36) = 1.09$] was significant. Because the game made no difference whatever to performance, we collapsed across this factor for the subsequent analyses. Notably, having all the feedback involved in the game and also having the partial feedback concerning the correctness of answers made no difference on any dependent measure.

Uncontrolled for lag to test. We conducted an analysis on the three feedback conditions uncontrolled for lag to test. The lag to test in the immediate feedback condition was 6.25 days ($SE = .21$), and in the delayed feedback condition, it was 2.77 days ($SE = .49$). As is shown in Figure 4A, there was an effect of feedback [$F(2,38) = 17.68$, $MS_e = .05, p < .0001, \eta_p^2 = .482$]. Delayed feedback gave rise to better performance ($M = .47$, $SE = .06$) than did immediate feedback ($M = .25$, $SE = .07$) or no feedback

(M = .07, SE = .03). All differences among these three means were significant, as determined by Tukey tests.

Controlled for lag to test. As in Experiment 1, we eliminated items given immediate feedback on Session 1 and given delayed feedback on Session 4, and then we conducted two analyses on the data controlling for lag to test. In the first analysis, we included the session on which the feedback had occurred (2 or 3) as a factor. As in Experiment 1, we had to eliminate many participants in this analysis. Only 12 participants had data in all three feedback conditions in each of the two sessions, and as a result, no effects were significant. In the second analysis, since session did not show any significant effects, we collapsed across sessions, but we computed the number of days between feedback and test in the immediate and delayed feedback conditions to be sure that collapsing did not introduce a lag-to-test confound, which it did not. The mean number of days from feedback to test in the collapsed data in the immediate condition was 4.67 days (SE = .43), whereas it was 4.12 days (SE = .32) in the delayed conditions. This difference was not significant [$t(19) = 1.15, p > .05$].

The main effect of feedback was significant [$F(2,36) = 3.91$, $MS_e = .06$, $p = .03$, $\eta_p^2 = .18$]. Means are shown in Figure 4B. The smallest difference between any of the three feedback means, as determined by a two-tailed Tukey test, was .18. The difference between the delayed feedback condition and the no-feedback condition was .18; the difference between the immediate feedback condition and the no-feedback condition was .18; the difference between the immediate feedback condition and the delayed feedback condition of .001 was not close to being significant.

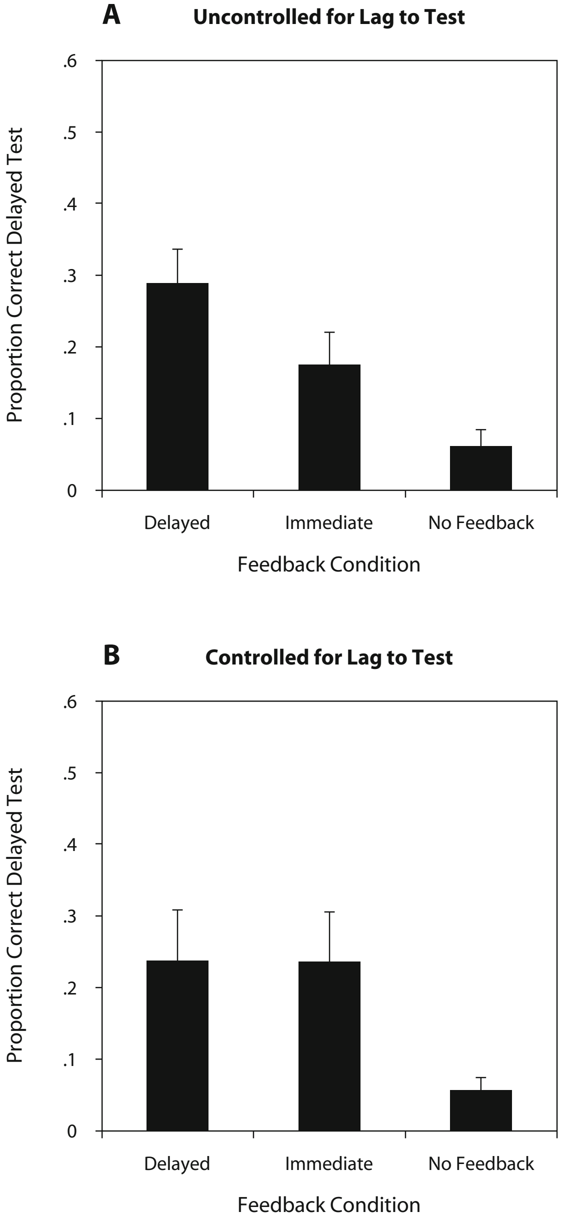

Figure 4. (A) Proportion correct as a function of feedback condition when lag to test was not controlled, with college students. (B) Proportion correct as a function of feedback condition when lag to test was controlled, with college students. Error bars represent standard errors.

GENERAL DISCUSSION

In both experiments, the difference between the delayed feedback condition and the immediate feedback condition—a difference that favored delayed feedback—was greater when the lag to test was uncontrolled than when it was controlled. Thus, both experiments implicate this lag-to-test factor as being important and as benefiting the delayed feedback condition. However, once lag to test was controlled, the results of the two experiments differed. The experiment with children still revealed a benefit for delayed feedback, whereas the experiment with adults showed no difference between the delayed and immediate feedback conditions.

Why, once this confound of lag to test was eliminated, might the college students not have benefited from delayed feedback, whereas the grade school children did? The first and most obvious reason is that adults and children might simply be different. Second, differential spacing effects have been proposed (Butler et al., 2007) as a reason that delayed feedback might be more beneficial than immediate feedback. However, there was little difference in the experiments in the objective amount of spacing between the immediate and delayed conditions for the children and that for the adults, since the experimental procedures and timing were similar for the two groups. Past research has shown that both groups benefit from spaced practice (Vlach, Sandhofer, & Kornell, 2008), but we could find no indication that the children benefitted more. Third, it had been thought (Kulik & Kulik, 1988) that studies conducted in the classroom might be different from those done in the laboratory. However, this possible difference goes in the wrong direction for our data. Past studies conducted in the classroom have favored immediate feedback over delayed feedback, whereas our experiment that was conducted in the classroom (Experiment 1) favored delayed feedback more than did the experiment conducted in the laboratory (Experiment 2).

Errors (of commission) were the fourth factor that had been proposed as being important in determining whether delayed or immediate feedback would be more beneficial. Accordingly, we looked for differences in this factor. As detailed in the introduction, however, there were two contrasting positions concerning the effects of errors. By Kulhavy and Anderson's (1972) view, errors should have been forgotten more when the feedback was given at a delay than when feedback was immediate. Less proactive inhibition would therefore occur in the delayed feedback condition, providing an advantage to that condition. By this view, the group with more errors should have benefited more from delayed feedback. The alternative view was that errors might be rehearsed, consolidated, and strengthened if they were not immediately corrected. Performance should have been enhanced by immediately eradicating errors, because they would then have less chance to harm memory for the correct answer. By this view, the group with more errors should benefit more from immediate feedback. By the first line of reasoning—given that the children benefited more from delay than did the adults—we expected to see more errors of commission in the children's data. By the second line of reasoning, we expected to see more errors of commission in the adults' data.

There was a difference in the proportion of errors of commission between the two experiments, and the adults made more of them (M = .61, SE = .07) than did the children (M = .40, SE = .06) [$t(42) = 2.22$, $p = .03$]. This difference between the two experiments and the finding that the children benefited more from delayed feedback suggest the possibility that when there are few errors to be contended with, delayed feedback is better. However, when there are many commission errors, the balance shifts toward immediate feedback. The finding that the benefit of delayed feedback was greater in the group with fewer, not more, errors goes against the idea of Kulhavy and Anderson (1972) that the benefit of delayed feedback results from decreased proactive inhibition resulting from the forgetting of errors. It seems, instead, that immediately correcting errors may be advantageous, but that there may be a tradeoff between the advantage and an opposing beneficial effect of spaced practice that occurs with delayed feedback.

The possibility exists, then, that it was the difference in the proportions of commission errors and whatever cognitive processes must be rallied to overcome these errors, rather than the differences in the ages of the participants, that led to the children showing a benefit of delayed feedback that the adults did not show. In the present research, we have sought to control one factor: lag to test. Our data indicate that this factor is important in determining whether immediate or delayed feedback is most helpful. When it is controlled, the benefits of delaying feedback are consistently diminished. But in what we thought was a simple age comparison that would produce equivalent results, we found an unexpected difference between children and adults. The finding that the number of commission errors was also different between these two groups leads to the conjecture that the proportion of commission errors may be an important determinant of when feedback should best be given. When there are many commission errors to correct, it may be less beneficial to delay feedback. These errors may, indeed, become entrenched during the delay, and it may be helpful to weed them out quickly. Experiments in which commission errors, the age of the participants, the delay of feedback (controlled for lag to test), and the spacing from the first presentation of the materials until the presentation of feedback are parametrically varied are needed to fully address this possibility. Such experiments will contribute to our understanding of how to best help children and adults correct their errors.

CONCLUSIONS

In most studies in which the relation between immediate and delayed feedback has been investigated, the time between the original test and the final test has been kept constant, but doing so results in the last presentation of the correct response, which is given by the feedback itself, being closer to the time of test in the delayed condition than in the immediate condition. When we allowed the feedback to be confounded with retention interval in this way, we, too, found that delayed feedback resulted in performance superior to those with immediate feedback. Although we treated this as a confound in our experiments, in practical terms, there is an obvious benefit to a short retention interval, which students would ignore to their peril. Having the last correct presentation of the correct answers as close to the test as is possible results in a test advantage.

In the experiments that we have presented here, one conducted with Grade 6 children and one conducted with college students, we were able to separate the effect of this often-seen retention interval difference between immediate and delayed feedback from other effects of delay of feedback by using a multiple-session design in which the retention interval of the delayed feedback and immediate feedback were equated. When we controlled retention interval, the recall data of the Grade 6 children still showed a benefit from delayed feedback relative to immediate feedback, but the recall data of college-aged adults did not. The difference might, of course, have been just a result of the age difference. However, the college students also made many more commission errors than did the children. We suggest—although we leave the definitive resolution to future research—that the difference in the beneficial effects of delayed feedback between the two groups might have been attributable to this difference in commission errors. Benefits in recall due to spacing effects (which are inherent to comparisons between immediate and delayed feedback) tend to favor delayed feedback. When there are few errors of commission, as in our children's data, the benefits of spacing weigh in strongly in favor of delaying feedback. However, when there are many errors, the benefits of spaced practice seen with delayed feedback may be offset by the recall advantage that accrues to correcting errors without delay.

AUTHOR NOTE

The research presented here was supported by CASL Grant R305H060161 from the Institute of Educational Sciences, Department of

Education, and by Grant 220020166 from the James S. McDonnell Foundation. The authors, however, are wholly responsible for the content of this article. We are grateful to Marc Meyer for his invaluable contributions to this research. We also thank Teal S. Eich and Noam Zerubavel for helping us to create Dragon Master. Correspondence concerning this article should be addressed to J. Metcalfe, 406 Schermerhorn Hall, Columbia University, New York, NY 10027 (e-mail: jm348@columbia.edu).

REFERENCES

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, **4**, 167-207. doi:10.1207/s15327809jls0402_2

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology*, **62**, 148-156. doi:10.1037/h0030766

Bjork, R. A., & Linn, M. C. (2006). The science of learning and the learning of science: Introducing desirable difficulties. *APS Observer*, **19**, 29.

Brackbill, Y., Bravos, A., & Starr, R. H. (1962). Delay improved retention on a difficult task. *Journal of Comparative & Physiological Psychology*, **55**, 947-952. doi:10.1037/h0041561

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

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

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

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. doi:10.1037/0033-2909.132.3.354

Guzman-Muñoz, F. J., & Johnson, A. (2007). Error feedback and the acquisition of geographical representations. *Applied Cognitive Psychology*, **22**, 979-995. doi:10.1002/acp.1410

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, **47**, 211-232. doi:10.2307/1170128

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, **63**, 505-512. doi:10.1037/h0033243

Kulik, J. A, & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, **58**, 79-97. doi:10.2307/1170349

Lhyle, K. G., & Kulhavy, R. W. (1987). Feedback processing and error correction. *Journal of Educational Psychology*, **79**, 320-322. doi:10.1037/0022-0663.79.3.320

Metcalfe, J. (2006). Principles of cognitive science in education. *APS Observer*, **19**, 29-39.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, **14**, 225-229.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology*, **19**, 743-768. doi:10.1080/09541440701326063

Murdock, B. B. (1974). *Human memory: Theory and data*. Potomac, MD: Erlbaum.

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. doi:10.1037/0278-7393.31.1.3

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

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, **16**, 183-186.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 601-604. doi:10.1037/0278-7393.30.3.601

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, **109**, 163-167. doi:10.1016/j.cognition.2008.07.013

APPENDIX
Materials Used for Experiment 2, With College Students

Amaranth	Imaginary flower reputed never to fade
Anemography	Treatise on the winds
Anomie	Condition of lacking accepted social values or standards
Aphesis	Loss of initial unaccented vowel from beginning of a word
Apiarian	Of, like, or pertaining to bees or beekeeping
Avuncular	Of or like an uncle
Badinage	Teasing conversation; banter; joking talk
Bivouac	Temporary encampment; camp without tents
Bloviate	To write or speak windily
Bromide	Platitude; chemical compound used to calm excitement
Calumny	Malicious misrepresentation; slander
Carillon	A set of bells (often in a tower) capable of being played
Chimerical	Fantastically improbable; highly unrealistic
Chine	Backbone and adjoining flesh of an animal
Codicil	Supplement to the body of a will; later addition to a will
Coeval	Living at the same time as; contemporary; of the same age
Complaisant	Trying to please; obliging
Connubial	Pertaining to marriage or the matrimonial state
Coquette	A flirt; woman who tries to attract men without sincere feelings
Cynosure	Object of general attention
Demotic	Of or pertaining to the people
Dilatory	Delaying; tending to delay

APPENDIX (Continued)

Doctrinaire	Uncompromising about principles; dogmatic; unyielding
Dyspeptic	Suffering from indigestion
Ebullient	Showing excitement; overflowing with enthusiasm
Effulgent	Shining brightly; brilliant
Emollient	Soothing or softening remedy (for the skin)
Emolument	Salary; payment for an office; compensation
Encomiastic	Praising; laudatory; eulogistic
Equable	Tranquil; of even temper; steady; uniform
Execrate	Curse; express abhorrence for; detest
Exiguous	Small in amount; minute
Heliotrope	A plant whose flowers turn to follow the sun
Hoary	Pale silver-grey color; grey with age
Immolate	Offer as a sacrifice; give up to destruction
Importunate	Always demanding; troublesomely urgent or persistent
Interstice	Narrow space between things
Larine	Of, like, or pertaining to gulls
Litotes	Understatement for emphasis
Mahout	One who rides or drives elephants
Manumit	Emancipate; free from slavery or bondage
Nembutsu	Buddhist invocation chanted to achieve enlightenment
Obloquy	Slander; disgrace; infamy
Obviate	Make unnecessary; get rid of
Pachyderm	Thick-skinned animal; an insensitive person
Palimpsest	Writing material used again after original text has been erased
Palliate	Ease pain (without curing); make less severe or offensive
Paroxysm	Fit or attack of pain, laughter, or rage; sudden outburst
Parvenu	Upstart; newly rich person
Patina	Green crust on old bronze or copper
Pellucid	Transparent; crystal clear; easy to understand
Penumbra	Partial shadow (in an eclipse)
Pertinacious	Holding tenaciously to an action; stubborn; persistent
Piebald	(Of an animal) consisting of two or more colors
Proscenium	Part of the stage in front of the curtain
Querulous	Given to complaining; fretful; whining
Quisling	Traitor who aids invaders; Benedict Arnold
Quotidian	Daily; commonplace; customary
Redolent	Odorous; fragrant; suggestive (of an odor)
Refulgent	Brilliant; brightly shining; gleaming
Reticulose	Of the nature of a network
Rostrum	Raised platform for speech making; pulpit
Sallow	Yellowish and unhealthy looking; sickly in color
Sidereal	Relating to or determined by the stars
Sinecure	Well-paid position with little responsibility
Stentorian	Extremely loud (of the voice)
Stipple	Paint or draw with dots or short strokes
Stygian	Unpleasantly dark; gloomy; hellish; deathly
Sybarite	Lover of luxury; person devoted to pleasure and luxury
Tremolo	Vibrating effect of certain musical instruments or the singing voice
Uranomania	Obsession with the idea of divinity
Uxorious	Excessively submissive or devoted to one’s wife
Voluble	Fluent; talkative; glib
Zoarium	Supporting structure for a polyp colony
Zwieback	Sweet toasted biscuit

(Manuscript received August 19, 2008;
revision accepted for publication June 30, 2009.)