2005

The Importance of Retrieval Failures to Long-term Retention: A Metacognitive Explanation of the Spacing Effect

Harry P. Bahrick
Ohio Wesleyan University

Lynda K. Hall
Franklin University, lynda.hall@franklin.edu

This is a pre-publication author manuscript. The published article is available at https://doi.org/10.1016/j.jml.2005.01.012.

Running head: THE IMPORTANCE OF RETRIEVAL FAILURES

The Importance of Retrieval Failures to Long Term Retention: A Metacognitive Explanation of the Spacing Effect

Harry P. Bahrick and Lynda K. Hall

Ohio Wesleyan University

Abstract

Encoding strategies vary in their duration of effectiveness, and individuals can best identify and modify strategies that yield effects of short duration on the basis of retrieval failures. Multiple study sessions with long inter-session intervals are better than massed training at providing discriminative feedback that identifies encoding strategies of short duration. We report two investigations in which long intervals between study sessions yield substantial benefits to long-term retention, at a cost of only moderately longer individual study sessions. When individuals monitor and control encoding over an extended period, targets yielding the largest number of retrieval failures contribute substantially to the spacing advantage. These findings are relevant to theory and to educators whose primary interest in memory pertains to long-term maintenance of knowledge.

The spacing effect is one of the oldest and best documented phenomena in the history of learning and memory research. Bruce and Bahrick (1992) found 321 publications on this topic, beginning with Ebbinghaus (1885/1913). The great majority of these investigations show that performance improves when practice is distributed rather than massed. However, the literature focuses neither on the effects of long intervals between practice sessions nor on the differential effects of spacing on acquisition versus long-term retention of content (Donovan & Radosevich, 1999). During the last thirty years, most investigations of the spacing effect have used tests of immediate retention and intervals of only a few seconds between repeated presentations of content (e.g., Hintzman, 1974). Current theoretical explanations of the spacing effect reflect these constraints.

In this paper, we present an explanation of the spacing effect that is focused on long-term access to knowledge in naturalistic learning situations. We provide an explanation based upon metacognitive monitoring of encoding strategies, and we present supporting data that document the benefits of widely spaced practice to long-term retention. We conclude with a discussion of the important contribution of retrieval failures to this phenomenon.

The Effect of Widely Spaced Practice on Acquisition versus Long-term Retention Early investigations of the spacing effect generally showed that learning was more rapid with shorter practice sessions and with longer intervals between sessions. Meta-analytic reviews (Donovan & Radosevich, 1999; Lee & Genovese, 1988) confirm the generality of these findings, particularly for motor tasks. In reviewing the early literature, Hovland (1951) reasoned that the benefits of increasing the length of the inter-session intervals during acquisition could only be obtained up to a maximum interval beyond which there would either be no additional benefits or an actual decrement of performance. Thus, when intervals between practice sessions reach a certain length, the forgetting that occurs between sessions impedes the learning process. Hovland's prediction was not explored, and research also failed to focus on the relation between performance during acquisition versus long-term retention.

We examined these relations in two investigations (Bahrick, 1979; Bahrick, Bahrick, Bahrick & Bahrick, 1993) of very long-term memory for foreign language vocabulary. Both investigations used training sessions with alternating study and test trials combined with a drop-out procedure so that words correctly recalled on a test trial were no longer studied or tested on subsequent trials. Training in each session ended with the first test trial on which all remaining words were correctly recalled. In the Bahrick (1979) investigation, participants studied English-Spanish word pairs for six training sessions with intertraining session intervals of 0, 1 day and 30 days. Figure 1 shows recall performance on training session 2 through 6 as well as on the final recall test. Performance on the first test trial of each training session was adversely affected by the longer intervals between sessions, i.e., there was more forgetting between successive sessions. However, when the interval between sessions six and seven was set at 30 days for all three groups, a cross-over interaction is apparent. Performance on the seventh session was superior for the group trained with the longest interval. The 30-day interval yielded continued improvement for the group trained with that interval, but it resulted in impaired performance for the two groups trained with shorter intervals. A follow-up retention test administered eight years later (Bahrick & Phelps, 1987) showed that participants trained with the 30-day interval still recalled 15% of the original word-pairs, those trained with the 1-day interval recalled 8%, and those trained with massed sessions recalled 6%. These findings were confirmed and extended in the second investigation (Bahrick, et al., 1993). In this study, four participants each learned 300 English-foreign language word pairs. The independent variables were the number of training sessions (13 or 26) and the inter-session interval (14, 28 or 56 days). In each of the 6 training conditions, participants studied 50 word pairs. With the longer intervals, training sessions required more study trials to reach criterion, but Figure 2 shows that the words acquired with the longer inter-session intervals are recalled better on retention tests administered one, two, three, or five years later. These results confirm Hovland's (1951) expectation that very long intervals between training sessions produce a reversal of the spacing effect; that is, more training time is

needed than with more closely spaced sessions to achieve comparable performance, but long term retention of content is enhanced.

It is important to note that the drop-out procedure used in the Bahrick (1979) and the Bahrick et al. (1993) investigations confounds the effect of spacing with the effect of the number of study and test trials. Although all word-pairs are correctly retrieved once in each session under all spacing conditions, wordpairs are likely to be studied and tested more often in the widely spaced condition prior to a correct retrieval, and this variable may affect long-term retention independently of the spacing effect. We will address the resolution of this confound later in this paper.

A Metacognitive Explanation of the Spacing Effect

Early explanations of the spacing effect were based upon work decrements associated with massed practice, or anti-consolidation, while current explanations focus on encoding variability (e.g., Glenberg, 1979; Madigan, 1969; Melton, 1970) or diminished processing (e.g., Bregman, 1967, Greeno, 1970; Hintzman, 1976) An important limitation common to prior theories is their relative neglect of the role played by conscious monitoring and control, and a corresponding emphasis on automatic, involuntary processing. Some investigators (such as Hintzman, 1974; Hintzman, Summers, Eki, & Moore 1975) concluded explicitly that voluntary attention is not involved in the diminished processing account of the spacing effect. However, a more recent focus on metacognitive research has led to a reconsideration of this issue. A number of investigators (Cuddy & Jacoby 1982; Elmes, Greener, & Wilkinson, 1972; Fishman, Keller, & Atkinson, 1968; Krug, Davis, & Glover, 1990; Pressley, Levin, & Ghatala, 1992; Shaughnessy, 1981; Shaughnessy & Zechmeister, 1992; Zechmeister & Shaughnessy, 1980; Zimmerman, 1975) report evidence that diminished processing may extend to longer spacing intervals and that conscious memory-monitoring mediates the differential encoding effects. However, most of these findings derive from investigations with repeated presentation of targets in a single list with short intervals between repeated presentations, and they show that participants process the later presentations of a target less than the earlier ones (e.g., Zimmermann, 1975).

The fact that learners use a variety of deliberate encoding strategies extensively in paired associate learning had been noted long ago (Reed (1918). More recently, Bellezza (1986) noted that encoding operations in associative learning constitute control processes on which learners can consciously reflect and therefore report verbally. In an excellent review of the mediation literature that included extensive new data, Richardson (1998) concluded: "Retrospective mediator reports provide valid accounts of the cognitive processes that occur at the time of learning and that play a causal role in determining the subsequent level of retention" (p. 597).

Our explanation of the long-term retention advantage of widely spaced practice sessions (Bahrick, 1979, 2000) assumes that encoding strategies vary in the duration of their effectiveness; that is, some strategies might facilitate target retrieval for only a few minutes, while others could remain effective for weeks or months. Further, we assume that learners can only discover that an encoding strategy has a short life-span by attempting to retrieve a target after their strategy fails. If retention is tested only at intervals shorter than the probable failure interval, the strategy remains effective, and the learner has no reason to replace or modify it. These assumptions can explain the cross-over interaction in Figure 1. Participants trained with short inter-session intervals had few opportunities to discover which of their encoding strategies were destined to have a limited duration of effectiveness. In the first 6 sessions, they failed to discover that many of their encoding strategies were inadequate to support longer term retention, and when a 30-day retention interval was applied for the first time prior to the seventh session, their overall retention declined sharply. Participants who were trained initially with 30-day intervals discovered during acquisition which of their encoding strategies were ineffective for long-term retention and could therefore replace or modify them. Each successive training session provided an additional opportunity to repeat this process. As a result, for participants trained with 30-day intervals, performance continued to improve on the first test-trial of the seventh session (Bahrick, 1979, 2000). We propose that the cumulative effect of exercising these metacognitive monitoring and control functions enhances long-term retention of content acquired under widely spaced training sessions.

Tests of the Metacognitive Monitoring Explanation

Experiment One

To test the assumptions outlined above, we examined learning and retention of 40 Swahili-English word-pairs by 11 male and 30 female Ohio Wesleyan University undergraduates who either earned course credit or were paid to participate in the investigation. We administered alternating study and test trials of the word-pairs, with a drop-out of word-pairs that were correctly recalled on a test trial. Training involved four sessions, and the between-group independent variable was the duration of the inter-session interval. We used intervals of 0 days (massed), 1 day, and 14 days, with 14 participants assigned to the massed and 14 day intervals and 13 participants assigned to the 1 day interval on the basis of their availability to each of the three schedules. All participants were given a final test trial 14 days after the fourth training session. Thus the total time involved was 57 days for the group trained with the 14 day inter-session interval, 18 days for the group trained with the 1 day interval, and 15 days for the group trained in the massed condition.

After each word-pair was presented for study, participants were required to report the strategy they had used for memorization. On subsequent presentations of the same pair, participants again reported their strategy as well as whether that strategy had been changed from the prior presentation of the same wordpair.

Procedure. Participants were tested individually or in small groups, with each individual seated in front of a personal computer. The procedure was guided by a program written in QBASIC by the experimenters. The experimenter told participants that the purpose of the investigation was to test their memory for Swahili-English word pairs and to find out more about how word-pairs are learned. The experimenter then described the following three common strategies of learning to associate word-pairs: a) repetition, b) verbal elaboration and c) visual elaboration, and gave examples of each. The experimenter gave examples to illustrate changing a strategy, for example, changing from repetition to a mediator, from a verbal to a visual mediator, as well as changing the details of a strategy, for example, a new verbal elaboration or a new visual image. Finally, participants were told that individuals differ in the strategies

that work best for them and that they should use whatever strategy they thought would be most effective.

The experimenter demonstrated the procedure by presenting five practice word-pairs that were not included among the 40 to-be-learned pairs (e.g., Fahali- Bull). Participants were seated in front of a computer, and the method of presenting and responding to individual word-pairs was explained. Each word-pair was presented for five seconds with the Swahili word appearing above the English word on the screen. An example of a Swahili-English word pair is Wingu-Cloud. After five seconds, the screen went blank, and the participants responded by indicating which of the following four methods they had employed to encode the word-pair: (a) simple repetition, (b) verbal elaboration by means of word or a sentence, (c) visual elaboration by means of a mental image, or (d) some other method. If participants had employed more than one method of encoding they responded by indicating all relevant categories. After recording their encoding method(s), participants pressed a key to elicit presentation of the next word-pair on a self-paced basis. A test trial followed in which each Swahili word was presented individually, and participants typed the associated English word on the keyboard. The test trials were also self-paced, and participants pressed a key to indicate that they were ready to be tested on the next pair. Responses were counted as correct if the first three letters matched those of the target English word, and participants were not provided feedback as to whether or not their response on each item was correct. Following the test trial, word-pairs that had not been correctly recalled were presented again in a new random order, and participants again indicated the method(s) used for encoding each pair. After they had answered the questions about the encoding method(s) employed, they also responded to a question asking if their encoding method had changed at all since the last time they tried to memorize that word-pair.

After the participants had mastered the five practice word-pairs, they were given an opportunity to ask questions and to clarify any ambiguities in the directions. The 40 pairs to be learned were then presented in a random sequence following the same procedure used with the practice pairs. A new random sequence was used on each study and each test trial, and the alternating study and test trials continued until the first test trial on which all remaining words were correctly recalled.

The procedure used in session 1 was followed again in sessions 2 through 4, with the exception that these later sessions started with a test trial, rather than a study trial. The first test trial included all 40 word-pairs. Participants assigned to the massed interval began a new session immediately after completing the prior session. The other participants returned to the laboratory either 1 or 14 days later, in accord with their assigned inter-session interval. All participants returned 14 days following the fourth study session for a final recall test.

Results and Conclusions. Figure 3 shows the proportion of correctly recalled targets on the first test trial for training sessions 2 through 4 and on the final retention test as a function of the inter-session interval. The three groups performed comparably on the first training session ($F(2,38) = 1.469$, $MSE = .016$, $p = .243$, $\eta^2 = .072$), but on sessions two to four, the massed group and the 1 day group performed considerably better than the group trained with the 14-day interval. Thus, the long inter-session interval adversely affected acquisition in that it yielded more forgetting between training sessions. However, the results of the retention test administered two weeks after the last training session again show the crossover interaction illustrated in Figure 1. The 14 day group continued to show improvement, while the performance of the other two groups declined. We evaluated the statistical significance of differences in percent recall among the three groups by an ANOVA in which the spacing interval was the between subjects variable and session the within subjects variable. The main effect of spacing interval was significant ($F(2,38) = 15.73$, $MSE=.085$, $p<,001$, $\eta^2 = .453$) as was the effect due to sessions ($F(3,114)=125.75$, $MSE= .009$, $p<.001$, $\eta^2 = .768$). The interaction effect of spacing interval with session was also significant ($F(6,114) = 93.91$, $MSE=.009$, $p< .001$, $\eta^2 = .832$). The interaction reflects the crossover of retention from session 4 to session 5. We conducted Bonferoni post hoc comparisons to evaluate the effect of intersession interval within each study session. For the second session, performance

in each intersession interval differed significantly from both of the others, with recall in the massed

condition

greater than recall in the day condition which was, in turn, greater than recall in the spaced condition. For the third and fourth study sessions, recall in the spaced condition was significantly below the other two conditions, which did not differ from one another. On the final recall test, performance in both spaced practice groups was significantly superior to performance of the massed group, but the performance of the two spaced groups did not differ. These results confirm that the effects of widely spaced training on acquisition are the inverse of its effects on long-term retention as noted by Bahrick (1979, 2000), Bjork (1994), Christina & Bjork (1991), and Schmidt and Bjork (1992). The widely spaced training intervals slowed acquisition but enhanced long-term retention.

Table 1 shows the mean number and percent of studied word-pairs on which participants reported using each type of strategy as a function of study session and intersession interval. As the percentage of correctly recalled targets increased from the first study session to the fourth, participants studied fewer word pairs and had fewer opportunities to report each type of study strategy. This trend was most pronounced with the massed interval. Participants in this condition studied fewer than four word pairs on average in the final study session. As a result, sample sizes were often too small to draw meaningful comparisons among the strategy types reported. To ease this problem, we identified word pairs on which the participants reported creating either a verbal or visual mediator (or both) at some point in the study session as well as word pairs on which they reported employing repetition exclusively. In the first training session, the spacing effect is not yet a variable. The massed training group reports using fewer mediators and more repetition than the other groups. All groups report using mediators (verbal or visual) more frequently than exclusive repetition. In subsequent training sessions, the 14 day group used mediators more than the other groups, and this is reflected in the total number of word pairs for which mediators were created and to a lesser extent in the percent of studied words on which mediators were used. By the fourth training session, only participants in the 14 day group reported frequent study of word pairs with either verbal or visual mediators. Correctly recalled targets were not presented again, and participants

in the massed and 1 day training groups who learned more quickly had fewer opportunities to re-encode these targets. However, when the use of mediators and of repetition are expressed as a percent of targets that are studied in each session, the 14 day group and the 1 day group continue to use mediators about twice as often as repetition, while the massed group uses repetition far more often than mediators. Participants in the 14 day group experienced more recall failures in sessions 2 to 4, and their failures provided more opportunities to re-encode targets using mediators. Their reports showed that they do just that. As a result, their retention of targets on the test administered two weeks later continues to improve significantly, while performance of the other two groups declines, giving rise to the cross-over interaction shown in Figure 3.

In subsequent analyses, we focus on the reports of creating new mediators under the three spacing conditions. Participants made these reports immediately after each word–pair was presented, and these data were reliable because they were based on short-term memory. In contrast, subjects' self reports of whether their strategies had changed or stayed the same were often in error, and we therefore chose not to analyze those data but instead recorded whether the strategy reported for an item was the same as or different from that reported when the item was studied in the previous session. We recorded all instances in which a subject who had not reported using a mediator for a word-pair in session one reported creating a mediator for that word-pair in one of the subsequent sessions.

The drop-out procedure controls the number of correct retrievals at 1 per target for all spacing conditions in each learning session, but as previously noted, the number of exposures/tests is generally greater under spaced practice. Better long-term retention under the spaced conditions could therefore reflect the larger number of exposures/tests rather than, or in addition to, the hypothesized differences in creating mediators in later sessions. In order to control the effects of the number of exposures/tests and to evaluate both the overall and unique contributions of spacing effects and mediator creation subsequent to session one, we subjected the data to a hierarchical multiple regression analysis. In this analysis, the

number of exposures/tests per word and the percent of word pairs with mediators created in session 1 (because this session was not affected by the spacing variable) were added simultaneously into the hierarchical regression as a control variable. The percent correct recall on the final retention test was the predicted variable, and the other predictor variables were two dummy variables created to represent the intersession intervals (one differentiated participants in the massed condition from the others; a second differentiated those in the 14 day condition from the others) and a variable representing mediator creation subsequent to session one for a word pair on which the participants did not report creating a mediator in the first session. Means and standard deviations for all variables included in the regression analysis are reported in Table 2, and intercorrelations among them are reported in Table 3.

Table 4 shows the results of the regression analysis when the mediator addition variable is entered either prior to or subsequent to the spacing variables. When the mediator addition variable is entered before the spacing variables, both steps in the hierarchy account for statistically significant ($p < .01$) increases in the accounted variance in final recall performance. However, when the order is reversed and the spacing variables are entered before the mediator addition variable, mediator addition no longer yields a significant improvement in predicting final retention. The regression analysis thus shows that whereas the spacing effects make a unique contribution to the prediction of final retention over and above the effects of the number of mediators created subsequent to session 1, the effect of the number of mediators created subsequent to session 1 makes no independent contribution to final retention beyond the spacing effect. Total final retention variance accounted for is a remarkable 85 percent.

The findings of experiment 1 reflect a situation in which the experimenters controlled the duration of exposure of the content, the number of tests and presentations of individual targets, as well as the criteria at which target presentations and test sessions were terminated. For these reasons it is unclear to what extent the observed, long-term retention advantages of spaced practice depend upon these experimenter-controlled parameters in contrast to more effective metacognitive monitoring and control

opportunities provided to the learner. In most naturalistic situations, such as a student preparing for a test, all of the above-mentioned parameters are controlled by the learner, not by an experimenter. We therefore designed the second investigation so as to approach naturalistic situations, that is, to have the learner control the above parameters. In this investigation, participants controlled exposure and encoding time, and the number of presentations/tests of all targets was held constant so that this variable could no longer contribute to any spacing advantage we might find.

Experiment two

Fifteen male and 22 female paid undergraduate students at Ohio Wesleyan University and 18 male and 20 female paid students at the University of Florida participated in the study.

Procedure. We trained participants using the15 most difficult Swahili-English word-pairs and the 15 easiest Swahili-English word-pairs from the Nelson & Dunlosky (1994) norms. Each training session began with a test in which the 30 Swahili words were presented individually in random sequence, and participants were asked to type the corresponding English word. In the first training session, this was done to ascertain whether participants were familiar with any of the Swahili words prior to the investigation, and two individuals who recalled several Swahili words were eliminated from the investigation on this basis. The 30 word-pairs were then presented individually on a personal computer in a random sequence with the English word appearing above the Swahili equivalent. Participants were instructed to study each pair until they had committed the pair to memory for a later recall test. They then pressed a key to solicit the next pair for study. The computer recorded the study time allocated to each pair. No feedback was provided on test trials. The interval between training sessions was the between subject variable. For the massed condition, new training sessions began immediately after the prior session ended; for the two spaced conditions sessions were administered on successive days and at 3 to 4 day intervals respectively. We allowed either three or four day intervals in order to avoid scheduling sessions on weekends.

Participants in all conditions were given five training sessions, with a single trial final retention test seven days after the last training session. We assigned participants to the three spacing conditions by taking into account their availability for being tested at various times, with 24 participants assigned to the massed schedule, 23 assigned to the 1 day schedule, and 26 assigned to the 3 to 4 day schedule.

Results and Discussion. Figure 4 shows recall performance for the three spacing conditions on the test trials preceding each study session as well as recall performance on the final retention test. Separate panels show these functions for the 15 easy and 15 difficult word pairs respectively. Longer intervals between practice sessions yielded a larger number of retrieval failures between study sessions, slowing down acquisition. The effect is observed for easy as well as for difficult targets, but it is somewhat more pronounced for difficult targets. Both panels show the cross-over interaction on the final retention test, with the two spacing conditions yielding superior retention to the massed condition. Performance continued to improve during the 7 day retention interval for the 3 to 4 day spacing condition, while performance did not change for the 1 day condition. We have shown (Bahrick, 1979; Bahrick et al. 1993) that the differential effect between the two spacing conditions becomes more pronounced if the final retention interval is longer in relation to the 1 day spacing interval, as can be seen in Figures 1 and 3 when longer final retention intervals were used.

We evaluated the statistical significance of differences in the number of correct responses among the three groups by means of an ANOVA with the spacing interval as the between subject variable and training sessions the within subject variable. The results confirm a significant main effect of the spacing interval ($F(2, 70) = 3.55$, $MSE=239.938$, $p< .05$, $\eta^2 = .092$), a significant main effect of the serial position of sessions ($F(4, 280) = 180.99$, $MSE =14.382$, $p< .001$, $\eta^2 =.721$), and a significant interval by session interaction ($F(8, 280) = 18.829$, $MSE = 14.382$, $p<.001$, $\eta^2 =.350$). Tukey post hoc tests indicate that the massed condition differed significantly ($p <.01$) from the 3 to 4 day condition in sessions 2 to 5; the

massed condition differed significantly ($p < .01$) from the 1 day condition in sessions 2 and 6, and the 1 day condition does not differ significantly ($p > .05$) from the 3 to 4 day condition in any session.

The data in table 5 show that study time is influenced far more by whether or not a target was retrieved on the prior test than by any other variable. The mean of every participant's median study time was 2.2 seconds (*SD* = 2.0) for targets that were retrieved on the prior test and 8.1 seconds (*SD* = 8.7) for targets that were not retrieved. The direction of this effect is consistent for all 73 participants in every study session ($p < .001$).

The critical inference from these data is that the larger number of retrieval failures associated with longer spacing intervals (shown in Figure 4) provided more differentiated feedback that improved the ability of participants to identify word-pairs they had encoded inadequately for long-term retention. Participants used this improved feedback effectively by increasing their subsequent study time. Based on the findings of experiment 1, we may assume that the increased study time was often used to create new mediators, but in other instances it may have been used only for additional repetition/rehearsals and contributed in that way to the total spacing effect.

The effect on study time is particularly pronounced for the more difficult targets, and it is obtained for errors of omission as well as errors of commission. Separate analyses based upon 375 errors show that 89% of errors are errors of omission, 11% are errors of commission, and mean study time subsequent to these two types of errors does not differ significantly ($p > .05$) .

Subsidiary analyses addressed the smaller effects of the serial position of encoding sessions and of spacing condition on the length of study time following either successful or unsuccessful retrievals. These effects are tangential to the major conclusion, and they must be interpreted cautiously. We performed two ANOVAS to evaluate the effect of spacing on study time following successful and unsuccessful retrievals respectively, with the data collapsed across acquisition sessions. The two spaced conditions produced somewhat longer study times than the massed condition following both successful and

unsuccessful retrievals, but neither effect is significant ($F(2,70) = 1.46$, $MSE = 4.006$, $\eta^2 = .04$ and $F(2,70) = .90$, $MSE = 76.646$, $\eta^2 = .025$, respectively).

Evaluating the effect of the serial position of study sessions on processing time by ANOVAS was problematic. The effects are confounded by a diminution of unsuccessful retrieval data on later learning sessions from participants who learned faster, and a corresponding diminution of data for correct retrievals on early practice sessions by participants who learned more slowly. We therefore evaluated the effects of serial position of the study sessions by means of sign tests comparing the duration of study for successive sessions for each participant. The results showed that study time for correctly retrieved targets diminished significantly ($p < .01$) and progressively on all successive study sessions. For incorrectly retrieved targets, study time decreased significantly ($p < .05$) between sessions 2 and 4, 2 and 5, and 3 and 5, but not between sessions 2 and 3, 3 and 4, or 4 and 5 ($p > .05$). The diminution of study time on later study sessions following incorrect retrievals may reflect the fact that word-pairs became progressively more familiar, and perceived familiarity may be interpreted to require diminished further study. These effects as well as all possible interactions must be interpreted with caution because successful vs. unsuccessful retrieval was not a variable manipulated by the investigators.

In order for memory research to become more relevant to education, investigations must approach certain naturalistic conditions. These conditions include longer, more realistic retention intervals as well as subject control of key parameters based upon metacognitive monitoring as illustrated in experiment 2. Accommodating these adjustments may require quasi-experimental (Campbell & Stanley, 1963) as well as correlational designs replacing traditional experimental designs, as well as corresponding departures from traditional ANOVA type analyses (Bahrick, 1994, 1996, 2005).

The Contribution of Retrieval Failures to Long-term Maintenance of Knowledge

Previous discussions of training conditions have stressed the value of successful retrieval practice. More specifically, Landauer and Bjork (1978) proposed that an expanding schedule of successful retrievals provides optimum conditions for the long-term retention of memory content, and their findings have been supported in subsequent investigations (Cull, 2000; Cull, Shaughnessy & Zechmeister, 1996; Rea & Modigliani, 1985; Siegel & Misselt, 1984). Although Landauer and Bjork explicitly limited their conclusions to conditions that provide no feedback for errors of omission or commission, this constraint has not been observed in the secondary literature. Thus Baddeley (1990) described expanding retrieval practice as a very powerful strategy that is easy to use, widely applicable, and probably more broadly useful than any of the more traditional visual imagery mnemonics (p. 158). Baddeley concluded that learners will be helped by being tested at a time when they can still remember an item; testing after the item is forgotten and then providing it is less conducive to learning.

We do not question the validity of conclusions regarding the benefits of successful retrievals or expanding retrieval schedules. Successful, spaced retrievals undoubtedly help to maintain access to memory targets based upon preventive maintenance (Bahrick & Hall, 1991), and the cumulative effects of successful retrievals permit increasing the intervals between retrievals without jeopardizing access. However, in practice, an expanding retrieval schedule that yields successful retrieval is feasible for an individual target item, or for a small number of targets, but much less practical for a larger number of targets. Targets vary in difficulty, and encoding strategies vary in their duration of effectiveness. Therefore, a retrieval interval that is appropriate for retrieving an easy target is likely to yield retrieval failures for difficult targets. The only options therefore are to use a retrieval interval short enough to yield successful retrieval of the most difficult targets or to vary retrieval intervals in accord with the level of difficulty of individual targets. Neither alternative is very practical when more than a few targets are involved.

The good news is that in situations where feedback is available, the value of retrieval failures followed by opportunities to re-encode inaccessible targets has been underestimated. Conservative assessments of the benefits of retrieval failures to long-term retention have been based on investigations that limit opportunities for metacognitive monitoring and control of encoding strategies. Others (e.g., King, Zechmeister, and Shaughnessy, 1980) have noted that knowledge of previous performance is assumed to be the basis of learners' decisions concerning encoding strategies on subsequent study trials and that conditions that adversely affect acquisition may benefit long-term retention (Bjork, 1994). Retrieval failures during acquisition are a case in point. Dempster (1989) concluded that spaced repetitions, regardless of whether they are in the form of additional study opportunities or successful tests, are a highly effective means of promoting learning, and Pashler, Zarow and Triplett (2003) confirmed the benefits of spacing, not withstanding a substantial increase in the number of prior retrieval failures under spaced practice. In both of our earlier investigations (Bahrick, 1979; Bahrick et al. 1993), the number of successful retrievals of targets was controlled at one per training session under all spacing conditions. It was the number of retrieval *failures* prior to a successful retrieval that varied as a function of target difficulty. In both investigations, *the spacing advantage in long-term recall is observed for word-pairs at all levels of difficulty, i.e., regardless of the number of retrieval failures and the most difficult words that produce the largest number of retrieval failures contribute heavily to the spacing advantage*. The results of the regression analysis in experiment 1 show that the spacing advantage for long term retention is independent of the variable number of tests/presentations associated with the drop-out procedure. In experiment 2 the number of target presentations is constant for all targets and all spacing conditions, and the data show that participants adjust study-time on the basis of their metacognitive monitoring of retrieval failures.

The feedback provided by retrieval failures becomes valuable only if learners are given an opportunity to adjust the duration or type of encoding. This process appears to require several repetitions.

Nelson and Leonesio (1988) found that people terminate study before learning is completed, even when they are instructed to master every item and allowed unlimited study time to do so. In agreement with Zacks (1969), with Metcalfe (2002), and with our own data, they report that learners allocate more study time to difficult items, but the adjustment is inadequate to achieve comparable recall performance. Nelson and Leonesio explain that learners overestimate mastery and give inadequate study time to difficult items because the access to most targets is from short-term memory at the time of study. Learners are better able to judge their true mastery on the basis of delayed recall tests. In the delayed condition, they must retrieve the information from long-term memory. Our findings confirm that longer intervals between practice sessions yield more retrieval failures and that this information helps learners identify their inadequate encoding. Such discriminative feedback from several delayed tests can explain the long-term advantages of the spacing effect. Naturalistic learning situations (e.g., preparation for an examination) can be arranged to benefit from this process. The preparation must involve several learning sessions spaced sufficiently to yield differential *failure* probabilities as a function of target difficulty. In previous research (Bahrick & Hall, 1991; Berger, Hall & Bahrick, 1999), we investigated the effectiveness of preventive and corrective interventions in maintaining/reinstating knowledge of long standing. Preventive interventions involve successful retrievals of old memory targets; corrective interventions involve a brief target presentation following retrieval failure. Both investigations showed that brief corrective interventions have very long-lasting effects on re-instating access to old targets. Our current results demonstrate that repeated, spaced corrective interventions help learners to identify inadequate encoding of newly acquired memory content and that this feedback yields long-term maintenance benefits superior to those obtained from preventive maintenance retrievals scheduled at short intervals. The latter are less useful because they fail to identify encoding strategies of short duration. We conclude that spaced retrieval failures preceding successful retrievals are more beneficial to long-term retention of difficult targets than an equal number of massed, successful retrievals.

Much recent research on the spacing effect has focused on single training sessions, short spacing intervals, short-term tests of retention and experimental control of target exposure. These conditions obscure the potential contribution of retrieval failures to long term access to knowledge. When learners are able to monitor and control their encoding during extended training, they make good use of the feedback provided by retrieval failures. In order to exploit these findings, educators will need further research to know how best to balance the cost of lengthened training with the benefits of extending longterm maintenance of knowledge.

References

Baddeley, A. (1990). Human memory: Theory and practice. Boston: Allyn and Bacon

Bahrick, H.P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. Journal of Experimental Psychology: General, 108, 296-308.

Bahrick, H.P. (1994). Extending the life-span of knowledge. In L. Penner, H. Knoff, G. Batsche & D. Nelson (Eds.), The challenge in mathematics and science education: Psychology's response. Washington, D.C.: American Psychological Association Press.

Bahrick, H.P. (1996). Synergistic strategies for memory research. In D. Herrmann, C. McEvoy, C. Hertzog, P. Hertel & M.K. Johnson (Eds.), Basic and Applied Memory Research, Volume 1, (pp. 5162). Mahwah, New Jersey: Erlbaum.

Bahrick, H.P. (2000). Long-term maintenance of knowledge. In. E. Tulving & F.I.C. Craik (Eds.), The Oxford handbook of memory (pp. 347-362). New York: Oxford University Press.

Bahrick, H.,P. (2005). The long-term neglect of long-term memory: Reasons and remedies. In (A.F. Healy (Ed.), Experimental cognitive psychology and its applications (pp. 89-100). Washington D.C.: American Psychological Association Press.

Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., & Bahrick, P.E. (1993). Maintenance of foreign-language vocabulary and the spacing effect. Psychological Science, 4, 316-321.

Bahrick, H.P. & Hall, L.K. (1991). Preventive and corrective maintenance of access to knowledge. Applied Cognitive Psychology, 5, 1-18.

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

Bellezza, F.S. (1986). Mental cues and verbal reports in learning. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 20, pp. 237-273). Orlando, FL: Academic Press.

Berger, S.A., Hall, L.K., & Bahrick, H.P. (1999). Stabilizing access to marginal and sub-marginal knowledge. Journal of Experimental Psychology: Applied, 5, 438-447.

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

Bregman, A.S. (1967). Distribution of practice and between-trials interference. Canadian Journal of Psychology, 21, 1-14.

Bruce, D. & Bahrick, H.P. (1992). Perceptions of past research. American Psychologist, 47, 319- 328.

Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin Co.

Christina, R.W. & Bjork, R.A. (1991). Optimizing long-term retention and transfer. In D, Druckman & R.A. Bjork (Eds.), In the mind's eye: Enhancing human performance (pp.23-56). Washington, DC: National Academy Press.

Cuddy, L. J., & Jacoby L. L. (1982). When forgetting helps memory: An analysis of repetition effects. Journal of Verbal Learning and Verbal Behavior, 21, 451-467.

Cull, W.L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. Applied Cognitive Psychology 14(3), 215-235.

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

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

Donovan, J.J., & Radosevich, D.J. (1999). A meta-analytic review of the distribution of practice effect:

Now you see it, now you don't. Journal of Applied Psychology, 84, 795-805.

Ebbinghaus, H. (1885/1913) Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Leipzig: Duncker und Humblot. (Translated as Memory: A contribution to experimental psychology, by H.A. Ruger and C.E. Bussenius. Columbia University College of Education. New York: Teachers College, Columbia University.)

Elmes, D.G., Greener, W.I., & Wilkinson, W.C. (1972). Free recall of items presented after massed- and distributed-practice items. American Journal of Psychology, 85, 237-240.

Fishman, E. J., Keller, L., & Atkinson, R.C. (1968). Massed versus distributed practice in computerized spelling drills. Journal of Educational Psychology, 59, 290-296.

Glenberg, A.M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. Memory and Cognition, 7, 95-112.

Greeno, J.G. (1970). Conservation of information-processing capacity in paired-associate memorization. Journal of Verbal Learning and Verbal Behavior, 9, 581-586.

Hintzman, D. (1974). Theoretical implications of the spacing effect. In R.L. Solso (Ed.) Theories in cognitive psychology: The Loyola Symposium (pp. 77-100). New York: Erlbaum.

Hintzman, D.L. (1976). Repetition and memory. In G.H. Bower (Ed.), The Psychology of learning and memory (Vol 11, pp.47-91) New York: Academic Press.

Hintzman, D.L., Summers, J.J., Eki, N.T., and Moore, M.D. (1975). Voluntary attention and the spacing effect. Memory and Cognition, 3, 576-580.

Hovland, C.I. (1951). Human learning and retention. In S.S. Stevens (Ed.), Handbook of experimental psychology (pp.613-689). New York: Wiley.

King, J.F., Zechmeister, E.B., & Shaughnessy, J.J. (1980). Judgments of knowing: The influence of retrieval practice. American Journal of Psychology, 93, 329-343.

Krug, D., Davis, T.B., & Glover, J.A. (1990). Massed versus distributed repeated reading- a case of forgetting helping recall. Journal of Educational Psychology, 82, 366-371.

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

Lee, T.D. & Genovese, E.D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered, Research, 59, 277-287.

Madigan, S.A. (1969). Intraserial repetition and coding processes in free recall. Journal of Verbal Learning and Verbal Behavior, 8, 828-835.

Melton, A.W. (1970). The situation with respect to the spacing of repetitions and memory. Journal of Verbal Learning and Verbal Behavior, 9, 596-606.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning. Journal of Experimental Psychology: General, 131, 349-363.

Nelson, T.O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. Memory, 2, 325-335.

Nelson, T.O., & Leonesio, R.J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." Journal of Experimental Psychology: Learning, Memory, & Cognition, 14, 676-686.

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

Pressley, M., Levin, J.R., & Ghatala, E.S. (1992). Memory strategy monitoring in adults and children. In T.O. Nelson (Ed.), Metacognition: Core readings (pp. 254-267). Boston: Allyn and Bacon.

Rea, C.O., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. Human Learning, 4, 11-18.

Reed, (1918). Associative aids: I. Their relation to learning, retention, and other associations. Psychological Review, 25, 128-155.

Richardson, J.T.E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. Psychonomic Bulletin and Review, 5, 597-614.

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

Shaughnessy, J.J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. Journal of Verbal Learning and Verbal Behavior, 20, 216-230.

Shaughnessy, J.J., & Zechmeister, E.B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. Bulletin of the Psychonomic Society, 30, 125-128.

Siegel, M.A., & Misselt, A.L. (1984). Adaptive feedback and review paradigm for computer-based drills. Journal of Educational Psychology, 76, 310-317.

Zacks, R.T. (1969). Invariance of total learning time under different conditions of practice. Journal of Experimental Psychology, 82, 441-447.

Zechmeister, E.B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. Bulletin of the Psychonomic Society, 15, 41-44.

Zimmerman, J. (1975). Free recall after self-paced study: A test of the attentional explanation of the spacing effect. American Journal of Psychology, 88, 277-291.

Author Note

This research was supported by grant number R01 AG19803-01 from the National Institute of Aging. We thank John Dunlosky for many helpful suggestions, Melinda Baker and Ann Daunic for supervising data collection and analyses, and the following individuals for collecting, recording, and analyzing data:

Kimberly Anderson, Laura Becker, Elizabeth Cremer, Jessica Junglas, Kevin Kula, Amanda Scott, and Kelly Weaver.

Correspondence concerning this article should be addressed to Harry P. Bahrick, Department of Psychology, Ohio Wesleyan University, Delaware, Ohio 43015

Table

Table 1.

Mean number of words and percent of studied words on which participants reported using each type of strategy as a function of study session and intersession interval.

	Number of Words				Percent of Studied Words			
	Study Session				Study Session			
	1	2	3	4	1	2	3	4
Repetition								
Only								
Massed	12.07	8.00	4.43	2.64	30.18	51.64	61.47	84.15
	(12.78)	(9.31)	(5.27)	(3.41)	(31.95)	(43.50)	(43.31)	(27.19)
1 day	6.77	5.92	2.54	1.23	16.92	25.22	41.25	48.89
	(7.53)	(7.74)	(3.07)	(1.74)	(18.82)	(28.42)	(42.34)	(48.77)
14 day	5.86	10.78	8.78	6.21	14.64	28.12	34.06	37.57
	(5.48)	(11.88)	(9.76)	(7.96)	(13.69)	(30.08)	(33.09)	(33.89)
Mediators								

Massed	24.86	4.00	1.00	.64	62.14	42.57	13.81	15.11
	(11.75)	(4.77)	(2.18)	(1.39)	(29.38)	(43.29)	(27.89)	(27.15)
1 day	30.00	14.08	4.31	2.31	75.00	67.87	57.47	51.11
	(10.49)	(10.08)	(6.84)	(5.15)	(26.22)	(34.11)	(41.23)	(48.77)
14 day	30.71	25.43	15.78	9.86	77.22	67.51	61.39	56.46
	(7.05)	(11.98)	(11.58)	(9.94)	(17.72)	(31.06)	(34.46)	(38.60)

Table 2.

Mean number of study exposures per word pair, percent of words with mediators in Session 1, percent of words with mediator additions in later study sessions, and percent correct on the final recall test as a function of intersession interval.

	Study Exposures	Percent Mediators in Session 1	Percent Mediator Additions	Final Recall
Massed	3.97	62.14	2.32	30.18
	(1.61)	(29.38)	(3.60)	(17.47)
1 day	4.08	75.00	4.81	73.46
	(1.92)	(26.22)	(4.94)	(13.75)
14 day	5.45	77.22	10.54	77.32
	(2.26)	(17.72)	(8.39)	(16.18)

Intercorrelations among dummy variables for spacing condition, mean number of study exposures per word, mediator variables, and percent correct on final recall.

Measure	1	2	3	4	5	6
1. Massed condition	--					
2. Spaced condition	-.519**	--				

Table 3.

3. Study exposures per word	-.194	.338*	--			
4. Percent mediators in Session 1	-.267	.169	-.178	--		
5. Percent mediator additions	-.383*	.492**	.018	-.139	--	
6. Final Recall	-.812**	.472**	-.228	.358*	.402**	--

* $p < .05$, ** $p < .01$

Summary of hierarchical regression analyses with percent correct on the final recall test as the criterion variable.

Step	Predictor Variables	R^2	ΔR^2	ΔF
			First Analysis	
1	Study exposures per word Percent mediators in Session 1	.156	.156	3.519*
2	Percent mediator additions	.363	.207	12.029***
3	Spacing variables	.847	.484	55.384***
			Second Analysis	
1	Study exposures per word Percent mediators in Session 1	.156	.156	3.519*

Table 4.

2	Spacing variables	.847	.691	81.133***
3	Percent mediator additions	.847	.000	.055

* ** ***
$p < .05$, $p < .01$, $p < .001$

Mean of median study time (in seconds) following successful and unsuccessful retrieval of targets.

	Study Session			
	2	3	4	5
Massed				
Successful	2.80 (2.59) n = 24	1.58 (1.04) n = 24	1.23 (.55) n = 24	1.13 (.50) n = 24
Unsuccessful	7.73 (6.46) n = 24	7.47 (10.11) n = 23	4.00 (2.75) n = 21	4.73 (3.05) n = 18
1 day				
Successful	5.99 (8.21) n = 19	2.89 (6.23) n = 23	1.39 (.83) n = 23	1.10 (.43) n = 23
Unsuccessful	9.33 (8.37) n = 23	10.22 (13.96) n = 23	10.60 (20.31) n = 21	6.32 (11.43) n = 17

Table 5.

3-4 day				
Successful				
	4.27	2.77	1.84	1.35
	(2.82) n = 19	(4.24) n = 26	(1.41) n = 26	(.65) n = 26
Unsuccessful	9.11	8.35	8.44	6.22
	(8.00) n = 26	(7.54) n = 26	(8.84) n = 25	(7.88) n = 23

Figure

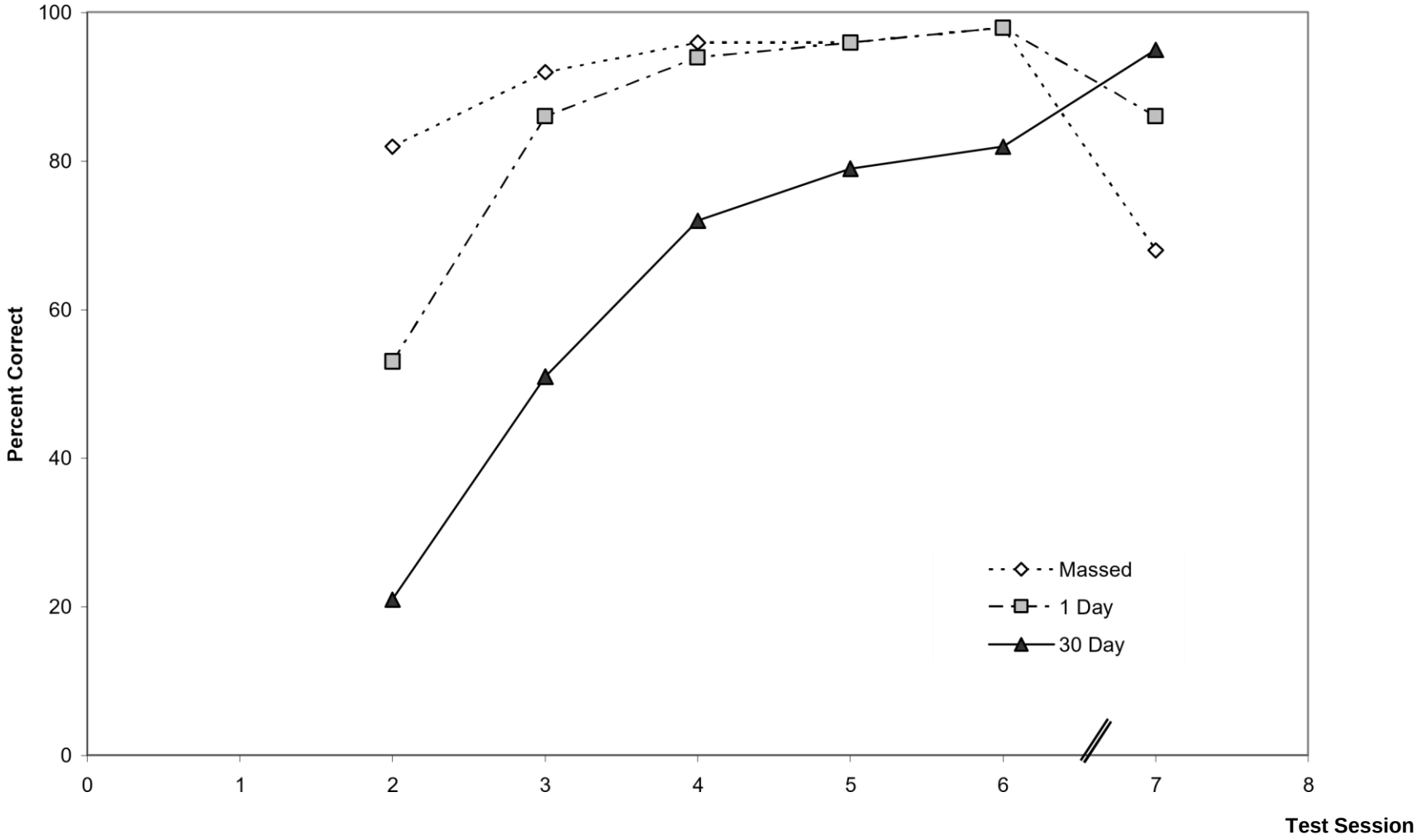
Figure Captions

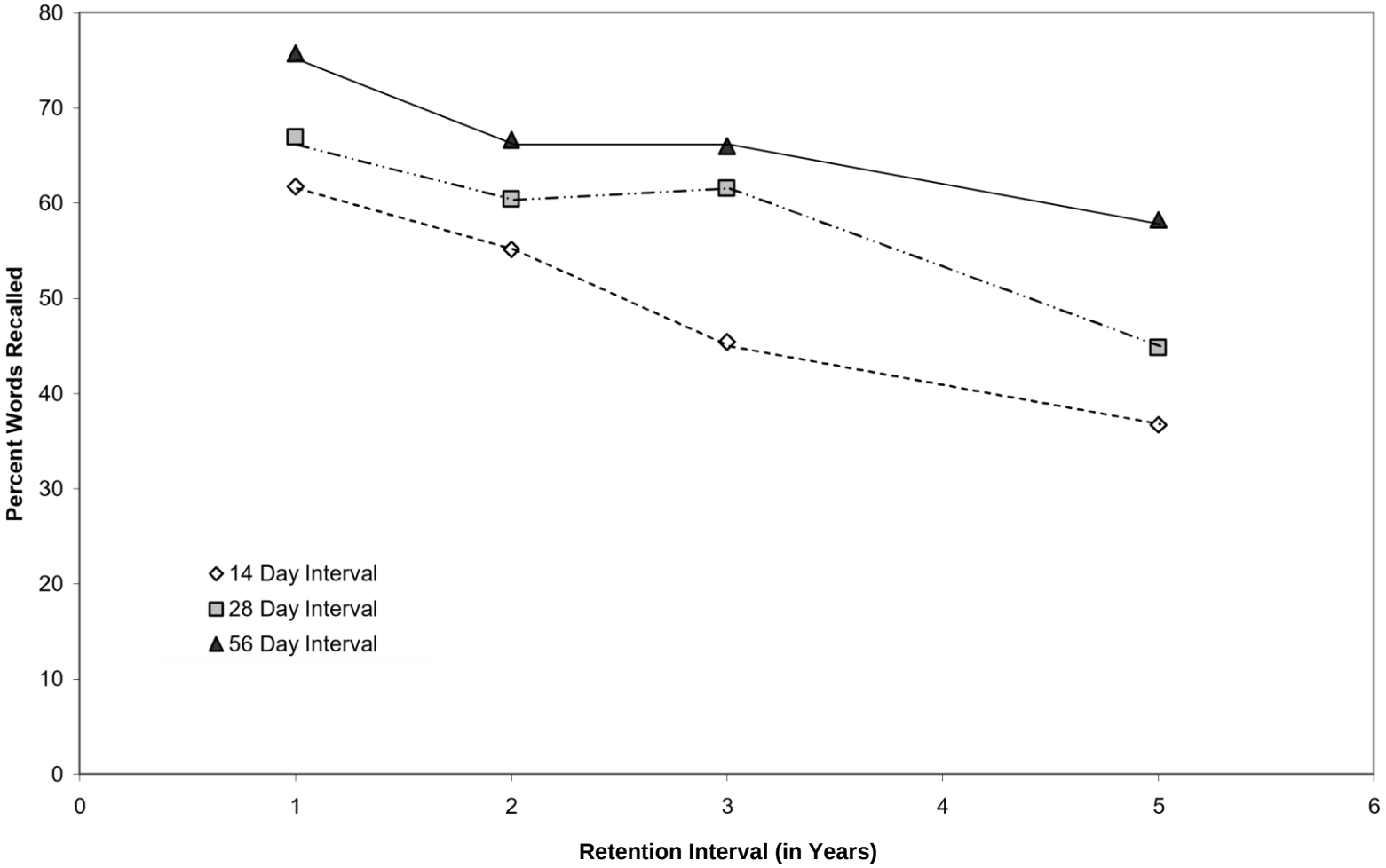
Figure 1. Percent correct on the first test trial as a function of test session and intersession interval (Bahrick, 1979).

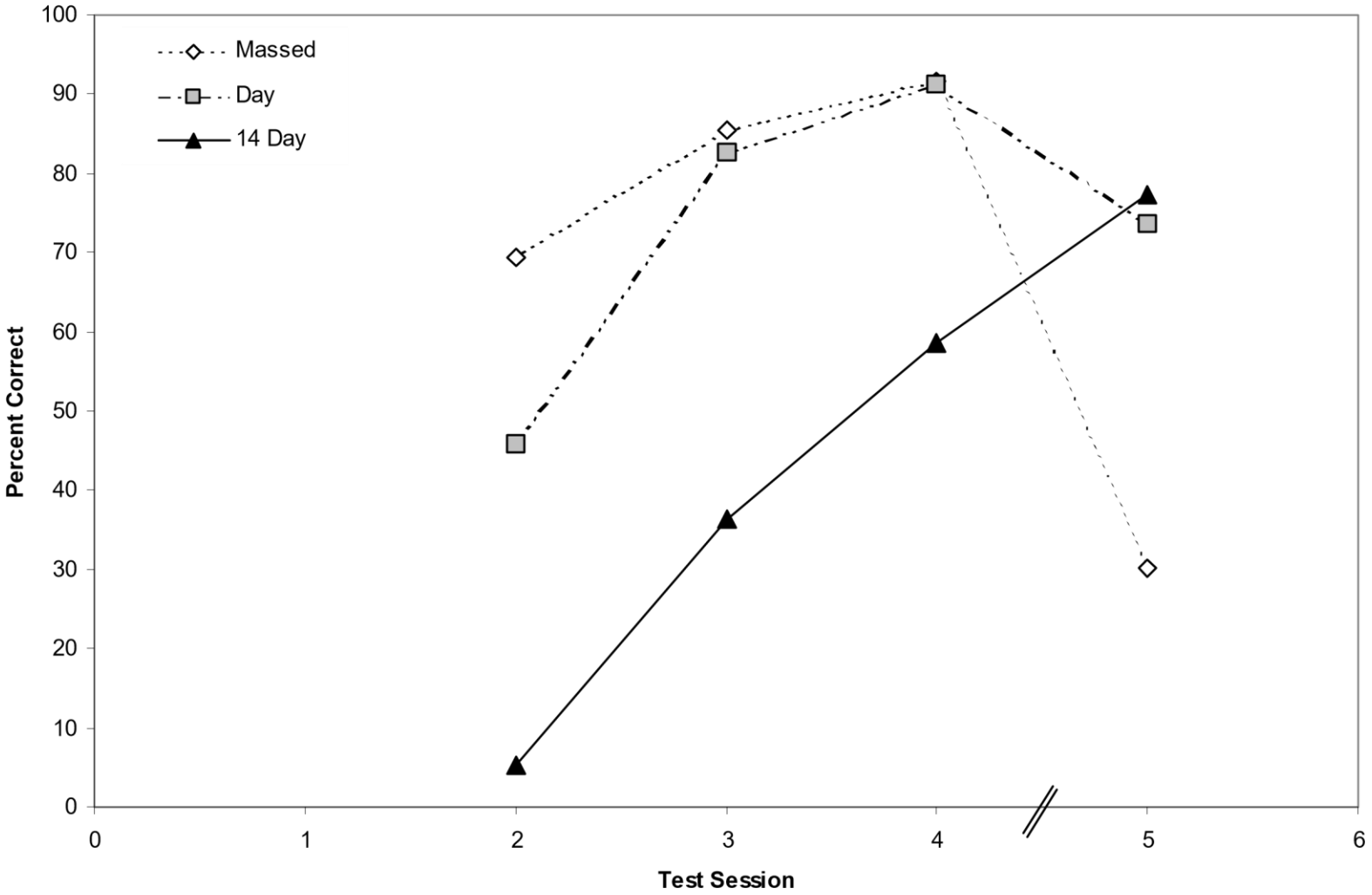
Figure 2. Percent of words recalled as a function of retention interval and intersession interval (Bahrick, Bahrick, Bahrick, & Bahrick, 1993).

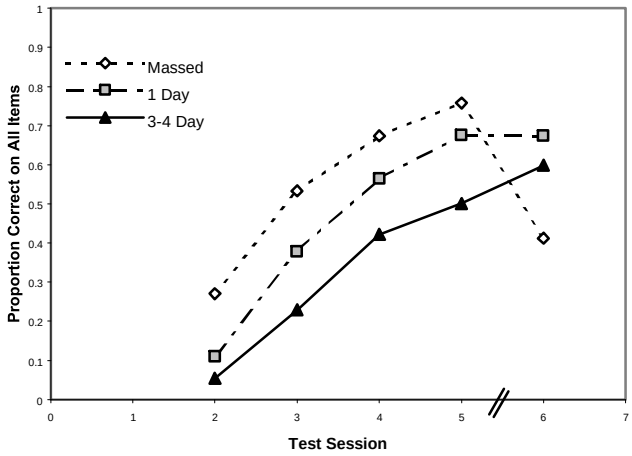
Figure 3. Percent correct as a function of test session and intersession interval.

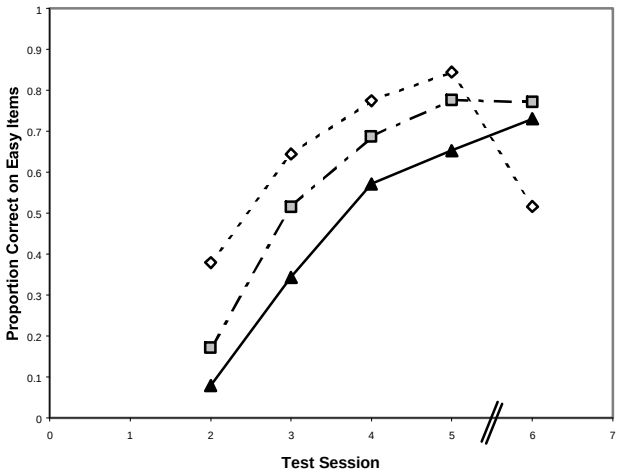
Figure 4. Proportion correct on all items, easy items and difficult items.

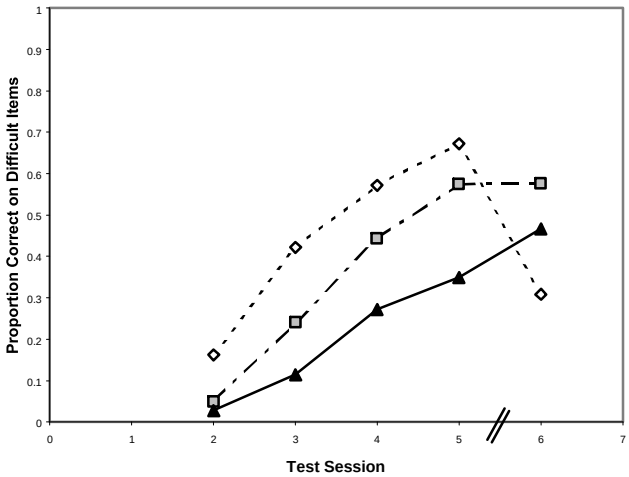

100
80
60
40
20
0
Percent Correct
0
1
2
3
4
5
6
7
8
Test Session
Massed
1 Day
30 Day


80
70
60
50
40
30
20
10
0
Percent Words Recalled
14 Day Interval
28 Day Interval
56 Day Interval
0
1
2
3
4
5
6
Retention Interval (in Years)


Massed
Day
14 Day
Percent Correct
Test Session
100
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6


Massed
1 Day
3-4 Day
Proportion Correct on All Items
Test Session


Proportion Correct on Easy Items
Test Session


Proportion Correct on Difficult Items
Test Session