Research Article

MAINTENANCE OF FOREIGN LANGUAGE VOCABULARY AND THE SPACING EFFECT

Harry P. Bahrick,[1] Lorraine E. Bahrick,[2] Audrey S. Bahrick,[3] and Phyllis E. Bahrick[1]

[1]*Ohio Wesleyan University*, [2]*Florida International University*, and [3]*University of Iowa*

Abstract—*In a 9-year longitudinal investigation, 4 subjects learned and relearned 300 English–foreign language word pairs. Either 13 or 26 relearning sessions were administered at intervals of 14, 28, or 56 days. Retention was tested for 1, 2, 3, or 5 years after training terminated. The longer intersession intervals slowed down acquisition slightly, but this disadvantage during training was offset by substantially higher retention. Thirteen retraining sessions spaced at 56 days yielded retention comparable to 26 sessions spaced at 14 days. The retention benefit due to additional sessions was independent of the benefit due to spacing, and both variables facilitated retention of words regardless of difficulty level and of the consistency of retrieval during training. The benefits of spaced retrieval practice to long-term maintenance of access to academic knowledge areas are discussed.*

The spacing effect was one of the earliest research topics investigated by experimental psychologists (Ebbinghaus, 1885/1964), and interest in the effect has sustained more than 300 investigations during the past century (Bruce & Bahrick, 1992). In spite of this considerable research effort, little is known about the long-term effects of spaced practice on the acquisition and retention of knowledge, and current explanations of the spacing effect (e.g., encoding variability theories and deficient-processing theories) may not be applicable to spacing effects obtained with long time intervals between training sessions.

Memory experiments require control of the conditions of acquisition and of rehearsals during the retention interval. Such controls are manageable in laboratory sessions lasting up to a few hours, but not when acquisition and retention extend over several years. The experimental paradigm is therefore not well suited to accommodate the long time periods during which knowledge systems are typically learned and maintained. To investigate the maintenance of knowledge over long time periods, it is necessary to accept diminished control over some conditions, or to use nonexperimental designs and statistical techniques to assess the effects of variables that could not be experimentally controlled.

An earlier cross-sectional investigation (Bahrick, 1984) showed that some of the knowledge acquired in Spanish language courses stabilized for at least 25 years under minimal rehearsal conditions, and that the share of original knowledge with such a long life span was a function of the number of Spanish language courses taken (i.e., it depended on the total amount and the distribution of practice). Because the amount and the distribution of practice were perfectly confounded in this naturalistic investigation (more courses involve more practice and also distribute practice over a longer period), follow-up investigations are needed in order to sort out the magnitude of these two effects and of their possible interactions. A quasi-experimental, longitudinal follow-up study (Bahrick & Phelps, 1987) found substantially enhanced performance on an 8-year retention test for subjects who learned 50 Spanish–English word pairs with a 30-day interval between successive relearning sessions in comparison with subjects who were trained for the same number of sessions with a 1-day intersession interval or who scheduled all the relearning sessions on the same day. The large recall benefit associated with spacing was achieved at a modest cost during acquisition; that is, subjects trained with a 30-day intersession interval required slightly longer acquisition sessions in order to reach comparable performance criteria. Table 1 shows the retention data reported by Bahrick and Phelps (1987).

The 1987 data show that increased spacing of training sessions greatly enhances long-term retention of vocabulary, with total practice held relatively constant. However, recall at the end of 8 years was quite low under all training conditions (overall, 9%), and the data do not show at what intersession interval the advantages of spacing would be lost, or to what extent the optimum interval depends on the number of training sessions or the difficulty of the material. The individual and interactive effects of these key variables must be understood in order to plan effective training programs, maintain effective rehearsal schedules, or develop models of memory that apply to the acquisition and maintenance of knowledge over long intervals. Current explanations of the spacing effect were developed to account for differential findings observed over short intervals, and neither encoding variability (e.g., Glenberg, 1979; Madigan, 1969; Melton, 1970) nor diminished-processing theories (e.g., Bregman, 1967; Cuddy & Jacoby, 1982; Greeno, 1970) are readily extended to spacing effects observed with very long intervals.

The purpose of the present study was to investigate the course of acquisition and retention of foreign language vocabulary as a function of the number of relearning sessions and the spacing between sessions. An additional goal was to relate the main and interactive effects of these variables to the difficulty of the material.

METHOD

Subjects

Recruitment of subjects was a challenge. The subjects' commitment extended over 9 years and required adherence to a

Address correspondence to Harry P. Bahrick, Department of Psychology, Ohio Wesleyan University, Delaware, OH 43015; e-mail: hpbahric@owucomcn.

Table 1. *Eight-year retention of foreign vocabulary as a function of the intersession interval (from Bahrick & Phelps, 1987)*

Primary intersession interval	Percentage correct		Percentage failed on both tests
	Recall test	Recognition test	
30 days	15	83	14
1 day	8	80	18
None	6	71	27
Control[a]	1	62	37

[a] Control subjects received no training.

demanding acquisition schedule, to self-monitoring of testing procedures, and to avoiding for 9 years extraexperimental exposure to the language from which vocabulary was being learned. The challenge was met by sharing investigative responsibilities and authorship among 4 subjects who are professional psychologists. Two of the subjects are older adults (age 57 at the beginning of the investigation) and 2 are young adults (ages 25 and 27 at the beginning of the investigation); 3 are female, and 1 is male.

Content Material

We determined that the vocabulary to be learned by each subject would belong to a foreign language in which the subject had prior training, but not recent training. This decision reflects the expectation that prior training in a language develops a language schema (Neisser, 1984); that this schema supports acquisition and retention of vocabulary on the basis of learned concepts, rules, and syntax; and that the schema increases the validity of metacognitive ratings and makes the learning situation more realistic in the sense that it resembles naturalistic language learning. We ruled out recent training in the target language because we wanted a knowledge base that was stable during the time span covered by the investigation (Bahrick, 1984, 1992). Based on these considerations, the language of choice was French for 3 of the subjects and German for the older female subject.

Vocabulary Selection

The prospective word pool for each subject was obtained from language texts and from a standard dictionary. Nouns, verbs (in the infinitive form), and adjectives were included in each pool. Each printed word in the pool was presented on an index card individually to the subject (by a participant not working with the same language), and the subject was asked the English meaning of the word. If the subject gave the correct meaning, the word was discarded. If the subject failed to give the correct meaning, he or she was asked to give a feeling-of-knowing rating, based upon whether he or she would be able to identify the correct English equivalent on a multiple-choice recognition test. If the rating was 1 (definitely yes) or 2 (probably yes), the word was discarded. If the rating was 3 (probably not), the word was retained. This selection procedure continued until 300 target words were retained for each subject.

Difficulty Ratings

Each of the target words was written on one side of an index card, with the English translation on the other side. Each subject was briefly shown each of the 300 word pairs and judged the ease of learning (EOL) the pair on a 3-point scale (Leonesio & Nelson, 1990). A rating of 3 designated word pairs the subject expected to be most difficult to learn.

Assignment of Words to Training Conditions

The 300 words for each subject were assigned to six training conditions (50 words to each condition). In order to control the expected difficulty level of words assigned to the various conditions, words with EOLs of 3 were assigned by systematic alternation among conditions. When this pool of words was exhausted, words rated 2, and then 1, were assigned in the same manner. The six training conditions orthogonally varied the number of retraining sessions (13 or 26) and the spacing between retraining sessions (14, 28, or 56 days) in a 2×3 factorial design.

Acquisition Procedure

Training sessions were self-administered following a preparatory session during which training conditions were discussed and rehearsed. The first training session began with a self-paced exposure trial. On this trial, the subject exposed each foreign word and after pronouncing the word, turned the card over to read the English equivalent. This exposure trial was followed by a self-paced test trial. The sequence of words for the test trial was altered by shuffling the index cards. The test trial consisted of exposing each foreign word and attempting to recall the English equivalent. If an English word was retrieved, the subject determined whether the response was correct by turning the card over. If the response was correct, the card was placed in a separate pile. If the response was incorrect, or if no response had been given, the card was placed in another pile. A second test trial immediately followed the first one. This trial was limited to the words that had been failed on the first test trial, but other procedures were unchanged. Each succeeding test trial was limited to the words failed on the preceding trial. The sequence of words was changed by shuffling the remaining cards between trials. The training session ended with the trial on which all remaining words were correctly retrieved. Thus, at the end of the training session, the subject had given one correct response to each of the 50 words, but correct retrieval had required a varying number of exposures for the 50 words. The subject recorded on each card the number of exposures required to retrieve that word.

Subjects began training with each set of 50 words on a different day, so that no more than one training session occurred

per day. Subsequent training sessions were scheduled according to the intersession interval appropriate for each set of words. Each of these subsequent sessions began with a test trial rather than an exposure trial; other aspects of the training procedure were unchanged from the first session. Training sessions continued for each set of 50 words in accord with the assigned schedule. Thus, training continued for 6 months under the 13-session, 14-day interval, and for nearly 4 years under the 26-session, 56-day interval schedule.

Retention Tests

The 50 words for each training schedule were assigned to one of four retention intervals (1, 2, 3, or 5 years) on the basis of their difficulty. The measure of training difficulty for each word was the cumulative number of exposures required for that word during all training sessions. The words were arranged in order of their difficulty level and were then assigned to the four retention intervals by a systematic alternation procedure of ABCD,DCBA,ABCD, and so forth. Thus, either 12 or 13 words were assigned to each of the four retention intervals.

Retention tests were performed 1, 2, 3, and 5 years after the last training session for each of the six training conditions. Each retention test began with a self-paced recall trial administered by an experimenter (usually a subject being trained in a different language). Words for which the correct English equivalent was recalled were separated out, and a five-alternative multiple-choice test was constructed by the experimenter for each of the remaining words. The four English foil words were selected from among English words of the same grammatical type as the target (noun, verb, or adjective), and they were words that the subject had learned as the correct responses to other targets. This was done so that familiarity with the English words in the training context could not provide a basis for subjects to differentiate targets from foils on the recognition test. Recognition tests for the words failed on the recall test were administered within a few minutes of the recall test, in the same testing session.[1]

RESULTS

Acquisition

Figure 1 shows the mean number of exposures per word per session, for the three intersession intervals. The data are averaged for the 4 subjects. The curves begin to level off after the seventh retraining session, but performance with the 56-day interval remains somewhat poorer throughout training. The difference does not reflect differences in the initial difficulty of the words assigned to the three intervals, as can be seen from performance on the first training session.

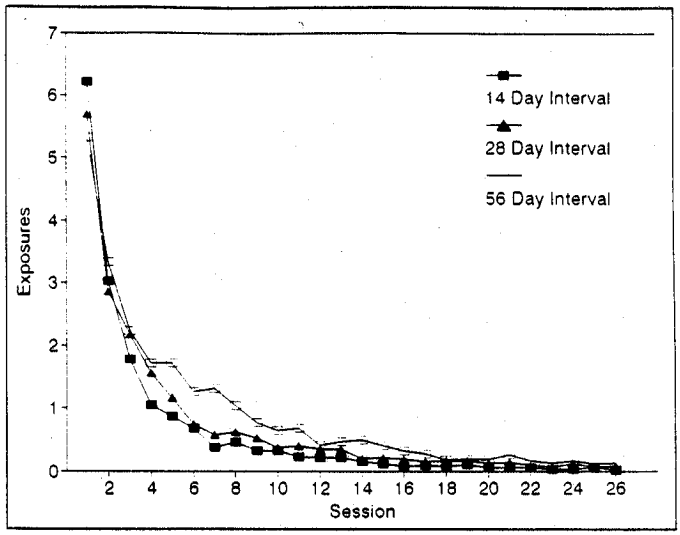

Fig. 1. Mean number of exposures per word per session during acquisition for the three intersession intervals.

Table 2 gives the mean number of words recalled on the first test trial at three stages of training. The data are presented separately for the 4 subjects. Two results are apparent from Figure 1 and Table 2: Not all word pairs are consistently retrieved even after 26 retraining sessions, and the number of words still requiring exposure at the end of training is somewhat larger for the longer intersession intervals than for the shorter intervals. The interaction showing diminishing effects of the training interval as the number of sessions increases is consistent for the 4 subjects.

The ceiling effect on the dependent variable renders the indicant insensitive to the retrieval practice that takes place during the later training sessions. The effect of these later sessions on long-term retention is very large, however, and it is discussed below.

Retention

Not all word pairs were mastered at the end of training. It is therefore appropriate to examine retention separately for those words that were recalled on the first test trial of the last training session. The top panel of Figure 2 shows these data as a function of the intersession interval. The data are combined for the 4 subjects and for 13 and 26 training sessions. At the end of training, the 14-day interval yields higher recall than the longer intervals, and this order is reflected in the first data point. The retention functions cross over during the 1st year of the retention interval; recall becomes lowest for the 14-day interval and highest for the 56-day interval, and this order is maintained throughout the 5 years.

The bottom panel of Figure 2 shows retention for all words,

1. Training and testing in this investigation were carried out by four investigators working in different parts of the world over a 9-year period, without the controls (e.g., of exposure, timing) that are standard in laboratory experiments. There were frequent departures from the scheduled intersession intervals, but the magnitude of these variations did not exceed 10% of the specified interval magnitude. More serious errors involved additional unscheduled training sessions and the omission of a terminal training session for one of the subjects, and a retention test scheduled for a 3-year interval that was administered after 4 years. These latter errors affect 2.14% of the acquisition data and 2.08% of the retention data. The erroneous data were omitted from Figure 2. There were also several failures to control extraexperimental exposure to vocabulary of the target language. Two subjects traveled in France for several days, and all of the subjects saw one or more films in their target language.

Table 2. *Mean number of words recalled on the first test trial of early, middle, and late training sessions, by subject and intersession interval*

	Subject 1			Subject 2			Subject 3			Subject 4		
Sessions	14 days	28 days	56 days	14 days	28 days	56 days	14 days	28 days	56 days	14 days	28 days	56 days
3–6	22	17	16	29	23	16	27	26	20	38	30	18
13–16	38	34	32	46	40	30	46	43	40	49	47	46
23–26	46	45	42	48	46	43	49	44	46	50	49	48

regardless of whether they were mastered at the end of training. The results confirm the superiority of more widely spaced training. Overall, the longer intervals yield substantially higher recall in spite of their adverse effects on acquisition. Table 3 shows the degree of consistency of these effects for the 4 subjects. The data are combined for 13 and 26 training sessions and for the 5-year retention interval.

When recall is examined for only those words that were failed on the first test trial of the last training session, recall (collapsed over 5 years) is much lower than for words that were retrieved (33% vs. 62%), but the differential effect of the intersession intervals remains very pronounced (22%, 29%, and 42%, respectively, for the three intervals).

We now examine interactions of the spacing effect with the number of training sessions, and with two indicants of word difficulty. Table 4 gives the mean percentage recall (collapsed over 5 years) as a function of the intersession interval and of the number of training sessions. Both variables have a major impact on retention, and there is no evidence of interaction. Thirteen sessions with a 56-day interval yield retention comparable to 26 sessions with a 14-day interval.

A criterion scale of consistency of retrieval of word pairs during the last five acquisition sessions yields additional information regarding the effects of the number of training sessions. Words were classified into three categories according to the number of times they were retrieved on the first test trial of the last five training sessions: never or once, two or three times, or four or five times. Analysis within each category reveals that long-term retention is consistently higher following 26 training sessions as compared with 13 training sessions, even when there is no ceiling effect and when the retrieval criterion at the end of training is controlled. Thus, the number of training sessions, the interval between training sessions, and the consistency of retrieval of individual words at the end of training independently affect recall over the 5-year retention interval, with no evidence of interaction among these variables.

We further examined the relation of word difficulty to retention when relative word difficulty was established individually for each subject. For this purpose, words were divided into upper, middle, and lower terciles based on the exposure distribution obtained from each subject for each of the three interses-

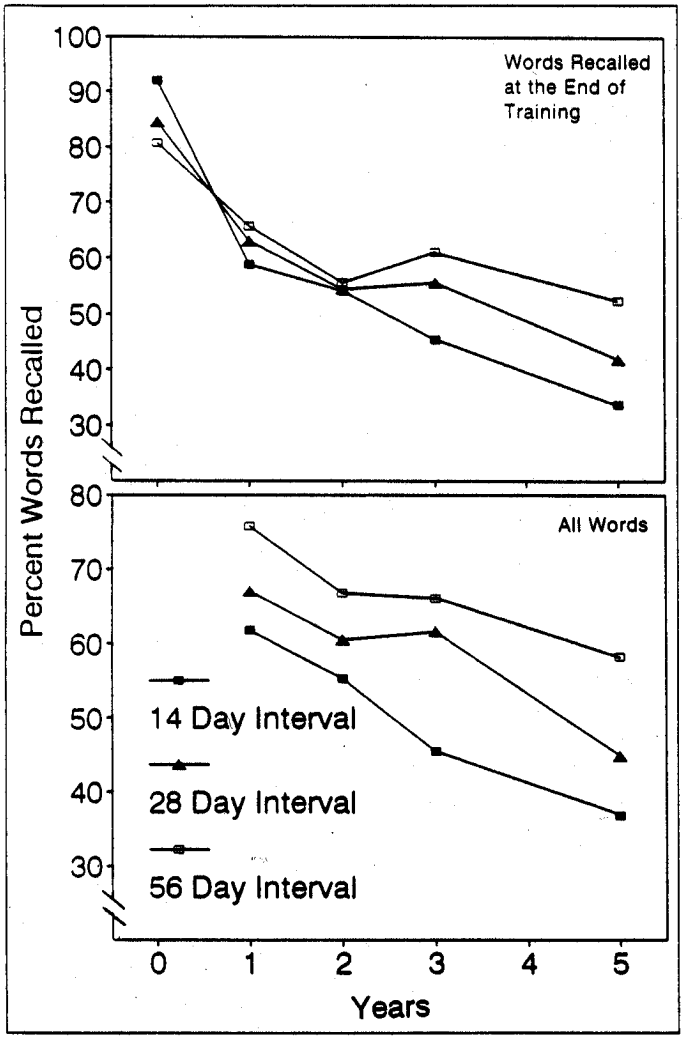

Fig. 2. Retention as a function of the intersession interval.

Table 3. *Percentage recall of individual subjects as a function of the intersession interval*

	Subject			
Interval	1	2	3	4
14 days	40	58	44	57
28 days	50	56	70	61
56 days	58	65	69	74

Table 4. *Mean percentage recall as a function of the intersession interval and the number of sessions*

Interval	13 sessions	26 sessions
14 days	43	56
28 days	50	68
56 days	57	76

sion intervals. The measure of difficulty was the cumulative number of exposures required for each word during the first 13 training sessions. Figure 3 shows that recall was highest for the 56-day interval and that this was true for words in all three difficulty terciles.[2]

DISCUSSION

Melton (1963, 1970) was among the first investigators to emphasize that learning curves in traditional verbal learning experiments show the net effects of learning and forgetting processes. This conceptual frame is appropriate for discussing acquisition in the present investigation, and it applies not only to successive trials within each acquisition session, but also to the cumulative effects of successive sessions spaced at various intervals. Recall increments across sessions reflect the net of increments within session and of decrements between sessions.

By the seventh retraining session, most, but not all, target words were recalled without requiring exposure. This finding agrees with the results obtained by Bahrick and Phelps (1987). Later training sessions involve primarily periodic retrieval. It is clear, however, that periodic retrieval has a very large effect on enhancing and stabilizing long-term retention. Recall after 8 years was 15% in the Bahrick and Phelps (1987) study based on seven retraining sessions and a 30-day intersession interval, and after 13 and 26 retraining sessions with a 28-day interval it was 40% and 54% after 5 years in the present investigation. The long-term effect of the additional training sessions was not limited to increasing the consistency of retrieval during training. Even when retrieval consistency at the end of training was held constant, increasing the number of training sessions improved long-term retention of words at various levels of retrieval consistency.

Increasing the intersession interval had two important consequences. During acquisition, the number of words that continued to require exposures in later training sessions was somewhat greater for the larger intervals. The number of such words was relatively small, and the differences between the intervals significantly diminished with training, but the effect was still evident at the end of training, as can be seen in Figure 1 and in Table 2. What is most noteworthy here is that varying the intersession interval from 1 day (Bahrick & Phelps, 1987) to 56 days has relatively minor effects on acquisition.

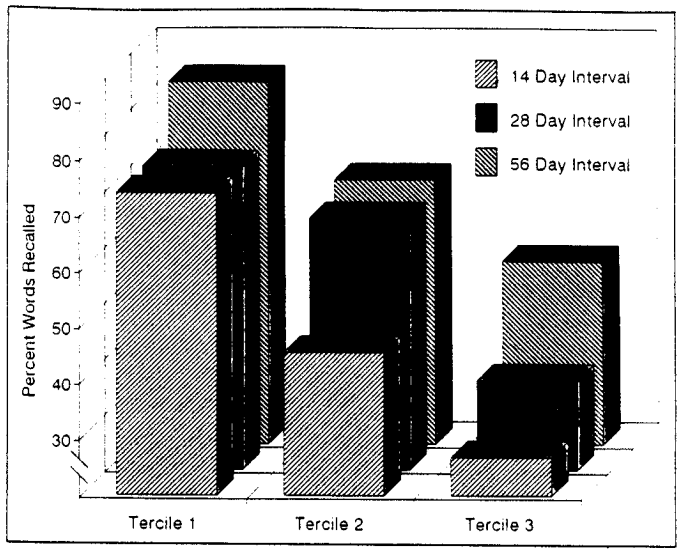

Fig. 3. Retention as a function of word difficulty (tercile of number of exposures during acquisition) and intersession interval.

The second consequence was the marked improvement of recall during the 5-year retention period associated with very long intersession intervals. The effect was not limited to those words that were mastered by the end of training, but was equally evident for words retrieved with various degrees of inconsistency.[3]

These findings illustrate dramatically one of the major points of a recent article by Schmidt and Bjork (1992): "Manipulations that maximize performance during training can be detrimental in the long term; conversely, manipulations that degrade the speed of acquisition can support the long-term goals of training" (p. 207). The explanation for these very large effects is not yet obvious because traditional theories of encoding variability and processing deficiency are based on time frames much shorter than those that affected performance here.

IMPLICATIONS FOR AN EDUCATIONAL STRATEGY

The life span of knowledge is an important, but neglected concern for educators and students. The cost-effectiveness of education depends upon how long knowledge acquired in schools will remain accessible, and research on maintenance of knowledge is needed to provide the relevant data. Other inves-

2. Other possible analyses are based on words failed on the recognition test or words failed on the recall test but passed on the recognition test. These analyses support the conclusions based on the recall test, but are less powerful because of a ceiling effect (recognition failures are limited to 6% of the data).

3. Because all 4 subjects had some knowledge of memory research, experimenter effects were more likely to occur than for naive subjects. All subjects expected longer intersession intervals to increase the difficulty of acquisition. However, none of the subjects held hypotheses in regard to which interval would yield overall optimum recall, and none of the subjects expected the length of the intersession interval to affect the recall of words that were not mastered at the end of training.

tigators (Dempster, 1988) have discussed the spacing effect as a prime example of the failure of educators to apply the results of psychological research. Although the literature overwhelmingly shows improved retention with distributed practice, the magnitude of the effects relevant to education cannot be estimated from the available laboratory data because the laboratory data are based on short periods of acquisition and retention. Relatively few investigators have tested long-term retention of academic content, and of those who have examined the spacing effect (Gay, 1973; Reynolds & Glaser, 1964; Smith & Rothkopf, 1984), none have used retention intervals longer than 31 days.

The present investigation shows that extended retrieval practice of foreign language vocabulary yields very large retention benefits over a 5-year period following the termination of practice, and that these benefits are greatest when the intervals between retrieval sessions are 2 months, or possibly longer. It is important to establish whether these benefits extend beyond the 5-year period under observation and whether they also apply to knowledge content that is more complex and integrative than individual word pairs. Previous findings based on cross-sectional data (Bahrick, 1984) are encouraging on both counts: For knowledge acquired in foreign language classes, very little, if any, additional forgetting occurs over the 25 years following the first 5 years. Also, the retention functions obtained for reading comprehension parallel those obtained for vocabulary retention. Reading comprehension involves the application of complex, schema-related knowledge to material that was not previously encountered; the finding therefore indicates that the obtained retention functions are not limited to the retention of simple associative connections.

Present curricula make few provisions for periodic retrieval practice of previously acquired knowledge. The powerful long-term effects of widely spaced retrieval practice constitute an important, but unexploited contribution of memory research to education. Optimally spaced retrievals offer the opportunity to extend the accessibility of knowledge at a cost that is low in relation to the overall costs of acquisition. Our present findings show that intervals of 2 months, or possibly longer, enhance long-term retention even for content that is not consistently retrieved at the end of training and for content that is more difficult to acquire under widely spaced practice conditions than under massed-practice conditions. Memory research must provide further data that reveal the costs and benefits associated with various retrieval schedules over time periods that are relevant to the interests of educators.

Acknowledgments—This investigation was supported by Grant DBS-9119800 from the National Science Foundation and Grant RO1HD25669 from the National Institute of Child Health and Human Development.

We thank Lynda Hall, Jeffrey Edwards, Krishna Tatenene, and John Wiebe for their important contributions to the data analysis.

REFERENCES

Bahrick, H.P. (1984). Semantic memory content in permastore: 50 years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113,* 1–29.

Bahrick, H.P. (1992). Stabilized memory of unrehearsed knowledge. *Journal of Experimental Psychology: General, 11,* 112–113.

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

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.

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.

Dempster, F.N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43,* 627–634.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H.A. Ruger & C.E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

Gay, L.R. (1973). Temporal position of retention of mathematical rules. *Journal of Educational Psychology, 64,* 171–182.

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

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

Leonesio, R.J., & Nelson, T.O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 464–470.

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. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior, 2,* 1–21.

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.

Neisser, U. (1984). Interpreting Harry Bahrick's discovery: What confers immunity against forgetting? *Journal of Experimental Psychology: General, 113,* 32–35.

Reynolds, J.H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology, 55,* 297–308.

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.

Smith, S.M., & Rothkopf, E.Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1,* 341–358.

(RECEIVED 9/21/92; REVISION ACCEPTED 12/23/92)