Cognitive Science 29 (2005) 559–586


# Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect

Philip I. Pavlik Jr., John R. Anderson

*Psychology Department, Carnegie Mellon University*

Received 4 September 2003; received in revised form 14 July 2004; accepted 17 November 2004

**Abstract**

An experiment was performed to investigate the effects of practice and spacing on retention of Japanese–English vocabulary paired associates. The relative benefit of spacing increased with increased practice and with longer retention intervals. Data were fitted with an activation-based memory model, which proposes that each time an item is practiced it receives an increment of strength but that these increments decay as a power function of time. The rate of decay for each presentation depended on the activation at the time of the presentation. This mechanism limits long-term benefits from further practice at higher levels of activation and produces the spacing effect and its observed interactions with practice and retention interval. The model was compared with another model of the spacing effect (Raaijmakers, 2003) and was fit to some results from the literature on spacing and memory.

*Keywords:* Spacing effect; Distributed practice; Memory; Forgetting; Practice; Mathematical modeling

## 1. Introduction

Although practice and forgetting have been researched extensively by psychologists for more than 100 years (Ebbinghaus, 1885), there is still no consensus on the mechanisms responsible for these effects. Central to finding this consensus is the need for explanations of how repetition improves recall, how increased temporal spacing of repetition improves recall, and how an increasing retention interval results in more forgetting.

Several theories exist. One major branch of theoretical explanation (Estes, 1955; Glenberg, 1979; Raaijmakers, 2003) explains the effects of practice and forgetting largely as due to *contextual fluctuation.* In this theory, each opportunity for practice results in an encoding of the stimulus and its context. However, the context of encoding fluctuates with the passage of time. Because of this fluctuation, as the spacing between repetitions increases, the overlap (redun-

Requests for reprints should be sent to Philip Pavlik, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: ppavlik@andrew.cmu.edu

dancy) of encoded contextual information decreases. This results in better memory performance as spacing becomes wider, because cue contextual information has a greater probability of matching the less redundant encoded information. Forgetting is also explained by contextual fluctuation theory because as retention intervals become longer the contextual information present in a retrieval cue has fluctuated to become more dissimilar to the encoding context, thus decreasing recall ability. Recently, Raaijmakers developed an effective mathematical model that realizes this type of theory. This article will take advantage of his work to compare the contextual fluctuation approach to the theory we will develop here.

Although versions of the contextual fluctuation theory are consistent with much of the data in the literature, other theoretical accounts are possible. Specifically, another school of thought (Cuddy & Jacoby, 1982; Schmidt & Bjork, 1992; Whitten & Bjork, 1977) says that the long-term memory strength contribution of a presentation depends on the accessibility of the memory at the time of the presentation. In this theory any repeated presentation that is more difficult (e.g., due to an impoverished stimulus or long spacing interval) results in a greater improvement in later recall ability due to this difficulty. Theories in this camp might loosely be referred to as accessibility theories because the accessibility of a presentation controls the long-run strength of the memory. One problem with these theories is that, in part because they have not been presented as fully specified formal models, it is unclear exactly how they would address all the effects in the literature.

As one instance of their lack of complete specification, accessibility theories have not made clear the details of how they would explain the crossover interaction between retention interval and spacing interval demonstrated by Bahrick (1979). Bahrick showed that when practice is spaced closely, it appears that forgetting occurs more quickly than when practice is spaced widely. He had subjects practice Spanish–English paired associates for six practice sessions separated by 1, 7, or 30 days and looked at retention 30 days after the final session. He found that final recall was significantly better as spacing between practice sessions was increased, even though the performance during practice was significantly worse with wider spacing.

Contextual fluctuation theories and models can capture these crossover effects using the interaction of contextual fluctuation and the redundancy of encoded traces. If the retention interval is short, closely spaced practices will be remembered better than widely spaced practices because the testing context will be similar to all the contexts of the closely spaced practices, but it will be similar to only the contexts of the most recent widely spaced practices. In contrast, if the retention interval is long, closely spaced practices will result in poorer recall because the test context will have fluctuated away from the overlapping encoding contexts, whereas widely spaced practices will result in better recall because the more diverse contextual information encoded will be more likely to match the test context.

We became interested in the issue of the exact nature of the accumulation of recall strength with spaced practice because we noticed in some pilot work that the Adaptive Character of Thought–Rational (ACT–R) 5.0 declarative memory equations were not fitting the data well. Specifically, in an experiment in which we intermixed different spacings of practice we were finding that we could not fit the data without supposing widely varying decay parameters for each condition. Because this pilot experiment was not designed to adjudicate these issues, we decided to design an experiment to address the issues of practice, forgetting, and spacing. Although verbal theories have provided interesting hypotheses of how these effects might occur, we hoped to use the data from this experiment to make a formal model of these effects because we felt modeling was necessary to address the complexity of the issues involved.

The formal model we were seeking needed to specify both a practice function and a retention function. Therefore, it is worthwhile reviewing the guidelines provided by Wickens (1999) for the necessary parts of a forgetting function, which can be generalized to practice functions. According to Wickens, these parts must include a representation for the initial learning, a description of asymptotic performance, and a way to characterize changes in the rate of forgetting.

First, the initial learning represents the strength of a memory item at the beginning of a retention interval. This quantity reflects the impact of study in the period preceding the interval represented in the forgetting function. From this initial level, memory strength decreases with time as forgetting occurs. Second, the forgetting function needs to explain performance after long retention intervals as it approaches an asymptote. According to Wickens (1999), asymptotic performance must be accounted for to explain results such as Bahrick (1984), which suggested that after about 3 years forgetting no longer occurs. Third, to characterize changes in the rate of forgetting, Wickens suggested calculating a "hazard function," which describes the rate of decrease of a memory at a particular time. He noted that the hazard function should agree with Jost's second law, which is essentially a statement that the rate of forgetting decreases with time: "If two associations are now of equal strength but of different ages, the older one will lose strength more slowly with the further passage of time" (Woodworth, 1938).

Although Wickens (1999) limited his analysis to forgetting functions, these aspects of forgetting functions match to similar aspects of practice functions. Practice functions also need an initial level, an asymptote, and a learning rate. However, unlike forgetting functions, which propose that forgetting is a continuous process, the learning rate in practice functions assumes some discrete increment for each added presentation. Therefore, our application of the learning rate must be framed as a combination rule that adds up the effects of separate presentations.

Our model will capture these effects with a strength function that has built into it both the power law of learning (a function of the number of practices) and the power law of forgetting (a function of the retention interval). There has been much debate over whether power functions are satisfactory for these purposes. Although some have argued that a power function characterizes practice data (e.g., Logan, 1992; Newell & Rosenbloom, 1981) and forgetting data (Wixted & Ebbesen, 1997), others have questioned whether a power function is satisfactory and have suggested that exponential functions are more suitable for practice data (Heathcote, Brown, & Mewhort, 2000). It has also been proposed that practice functions may be a mixture of multiple power functions (Delaney, Reder, Staszewski, & Ritter, 1998; Rickard, 1997). Still others have argued for the superiority of an exponential-power function in forgetting (Rubin & Wenzel, 1996; Wickelgren, 1974) and practice (Heathcote et al., 2000).

Our choice of a power function was based on analyses of how need probabilities for memories fluctuate in the environment (J. R. Anderson & Schooler, 1991), yet it remains unclear why need probabilities for memories should follow power functions. Because it has been shown that a mixture of exponential processes can produce power-law-like functions (R. B. Anderson & Tweney, 1997), one might speculate that a mixture of exponential decays in need probability in the environment is the cause. Regardless, because biological processes often follow exponential decay functions, it is not difficult to suppose that forgetting matches power functions in the environment because of a mixture of exponential processes in the brain. Essentially, the functional form we propose was chosen because it summarizes all the major effects of relevance with a tolerable degree of error.

Our explanation of the integration of practice and forgetting began with this ACT–R memory model (J. R. Anderson & Lebiere, 1998). This model specified how the exact pattern of practice and retention mapped to memory performance. However, it did not explain how these factors might interact with the spacing of practice. The model assumed that each presentation results in an increment to memory, that these increments decay according to a power function, and that these decaying increments sum to yield an overall strength of the memory trace. Using this model, J. R. Anderson, Fincham, and Douglass (1999) reported success in fitting various latency data, but for this article, we modeled correctness of recall to facilitate comparisons with other work.

To study practice and forgetting, we chose a paired-associate memory task in which participants memorized the English translations of Japanese words. Foreign language vocabulary learning involves a basic memory task, but it still has external validity, and thus it seemed to be an ideal paradigm. Japanese was chosen to minimize the prior learning participants could bring to the task.

## 2. Experiment

### *2.1. Participants and design*

Forty participants were recruited for this study from the Pittsburgh, Pennsylvania, community. They were mostly college students responding to an online advertisement. All participants completed the experiment. Twenty participants each were assigned to the 1- and 7-day retention conditions. Sessions lasted between 60 and 90 min. Only participants who professed no knowledge of Japanese were recruited.

In this experiment, participants learned the Japanese–English paired associates during a first session (S1), and then 1 or 7 days later returned for a second session (S2) to assess their retention. During S1, participants learned the English responses for the Japanese cues over the course of 12 blocks of 40 presentations each. A presentation consisted of either a study trial or a test trial with feedback. The first 26 presentations of the first block were buffers and were not analyzed. Following these buffers, the word pairs for each condition were introduced with study trials and then tested 1, 2, 4, or 8 times with 2, 14, or 98 intervening presentations. This indicates a $3 \times 4$ design (not including the two levels for the between-subject long-term retention factor); however, because it would have made the experiment take too long, the $8 \times 98$ condition was not included, resulting in 11 within-subjects conditions. Each condition used eight word pairs. The introduction of pairs in each condition was distributed across the span of S1. Buffer items were used to fill in presentation spaces in the blocks that were not needed for the 11 conditions.[1]

On S2, after a 1- or 7-day retention interval, the effects of these 11 conditions of training were assessed with 11 blocks of 40 test trials. There were 27 buffer items to begin the first block of S2. Following these buffers, all word pairs were retested four times each at a spacing of 98 presentations between retests.

### 2.2. Materials

The stimuli and buffers were 104 Japanese–English word pairs. English words were chosen from the MRC Psycholinguistic database such that the words had familiarity ratings between 406 and 621, with a mean of 548, and had imagability ratings between 343 and 566, with a mean of 464. These ratings were composed according to procedures described in the Medical Research Council (MRC) Psycholinguistic Database manual (Coltheart, 1981). The overall MRC database means for familiarity and imagability are 488 (*SD* 120) and 438 (*SD* 99) respectively, so the words we chose had higher familiarity and imagability ratings than the database averages. Japanese translations (from the possible Japanese synonyms) were chosen to avoid similarity to common English words. Only four-letter English words were used, and four- to seven-letter Japanese translations were used. Japanese words were presented using English characters. Word assignment to conditions was randomized for each participant.

### 2.3. Procedure

Participants were instructed not to practice the word pairs during the time between sessions. Participants were scored for motivational purposes, receiving 6 points for each correct response and losing 12 points for each wrong response. Failing to provide a response, either by time-out or by providing a blank response, resulted in a 0 score. Participants were paid $9 to $15 per session depending on their score.

The stimuli were shown on a 19-in. monitor at a resolution of 1,024 × 768 in 48-point white Tahoma font on a blue screen. The stimuli pairs were centered vertically, the Japanese words appearing on the left and the English words on the right side of the screen. Participant prompts appeared centered horizontally, slightly above the words. Participant prompts were in 37-point Tahoma.

All trials were cued with the prompts "Study" or "Test" for 2 sec. Study opportunities allowed participants to view the new pair for 5 sec. Tests involved presentation of the Japanese word on the left side of the screen. Participants typed the English translation on the right. If no response was made, the program timed out in 7 sec. Following response or failure to respond, the program displayed "Correct" or "Incorrect" for 1 sec and showed the change of score. If the response was correct, the next trial began. If incorrect, the word "Restudy" appeared for 2 sec, and there was a 5-sec restudy opportunity, identical with the original study trial.[2]

Between the blocks of 40 items, participants continued by pressing the space bar when they were ready. Few participants paused at these opportunities. S2 procedures were identical, with restudy trials after incorrect responses, but no new words were introduced.

The "recall or restudy" procedure for presentation of the stimuli was chosen based on the assumptions of the model we would be fitting to the data. According to this ACT–R memory model, study presentations and successful test presentations benefit memory equally. Therefore, our procedure results in one memory increment (according to the model) for each trial regardless of whether the participant responded correctly.

Although our model assumes each study or test counts equally, Carrier and Pashler (1992) noted in their literature review that studies and tests are not exactly equal. The identical credit we give to test and study trials is essentially a simplifying assumption to reduce model complexity so the spacing effect model can be studied independently. Carrier and Pashler showed the differences between studies and tests are not always large, so considering them equal is a reasonable approximation for simplifying our model.

## 3. Results and discussion

By spacing the introduction of pairs across the session, S1 was designed to avoid confounding the conditions with serial position effects. Because of this, it did not seem that there should be significant differences between correctness on the first two trials for items that were to receive two practices, those that were to receive four, and those that were to receive eight, and indeed there was none. This was shown through two repeated measures ANOVAs that were completed to look for differences in the means of these first two trials across the different practice conditions. To deal with the fact that the eight-repetition, 98-spacing condition was missing from a complete factorial design, we performed two analyses of variance (ANOVAs) on subsets of the design that were fully factorial. The first ANOVA was performed to compare the mean percent correct on the first two presentations for two and four practice conditions (0.542 and 0.515, respectively), averaged over 2, 14, and 98 spacing conditions. The difference was not significant, $F(1, 39) = 3.1$, $p > .05$. The second ANOVA was performed to look at the means for two, four, and eight practices (0.662, 0.643, and 0.653, respectively) aggregated for 2 and 14 spacing. There were no significant differences, $F(2, 38) = 0.60$, $p > .05$. We also looked for serial position effects within conditions. The eight items of each condition were introduced in sequence across the experiment, so we looked at the first test trials of all items and found averages of correctness for these first trials by their order of introduction. This gave eight means spread fairly evenly across the experiment. An ANOVA looking for differences in these means found nothing significant, $F(7, 273) = 1.925$, $p > .05$.

Because of the similarity across practice conditions, S1 data were aggregated for display (see Fig. 1). To confirm that performance improved across this first session and to look for effects of the spacing manipulation during S1, a repeated measures ANOVA (first four trials of S1 aggregated by Practice condition × S1 spacing × S2 retention interval group) was completed. The results of this analysis confirmed that performance improved across the first four trials of S1, $F(3, 114) = 412$, $p < .001$, and that there was significantly lower performance with wider spacing, $F(2, 76) = 240$, $p < .001$. The lower performance on S1 with wider spacing was likely due to the overall longer retention intervals for these trials. The difference in S1 learning for the 1- and 7-day retention groups was not significant on S1, $F(1, 38) = 1.13$, $p = .296$.

Next we looked for a crossover spacing interaction over the retention interval by comparing the means for the last trials on S1 with the means for the first trials on S2, using a repeated measures ANOVA (Trial × S1 spacing condition), excluding the eight repetition conditions. The data can be seen in Fig. 2. The interaction of Trial × S1 spacing condition was strong, $F(2, 76) = 321$, $p < .001$. This interaction suggested presentation sequences with wider spacing resulted in less forgetting.

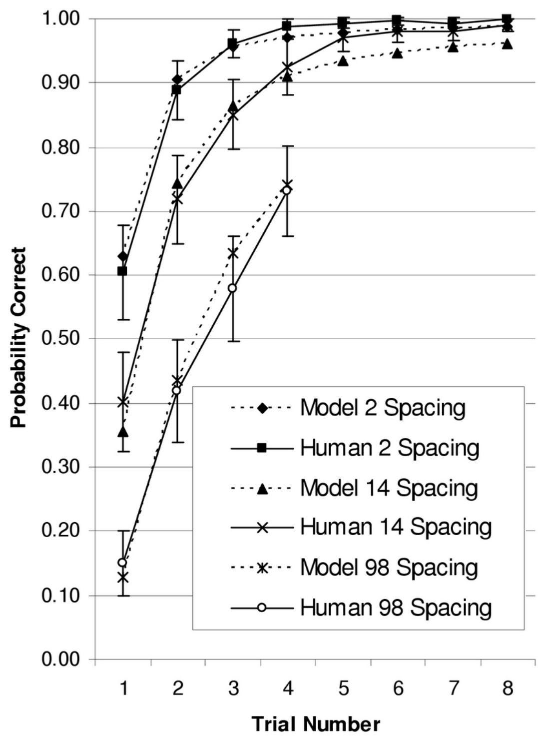


Fig. 1. Experiment S1 aggregate data for humans and model for spacing conditions. 2 *SE* confidence intervals computed from participant means.

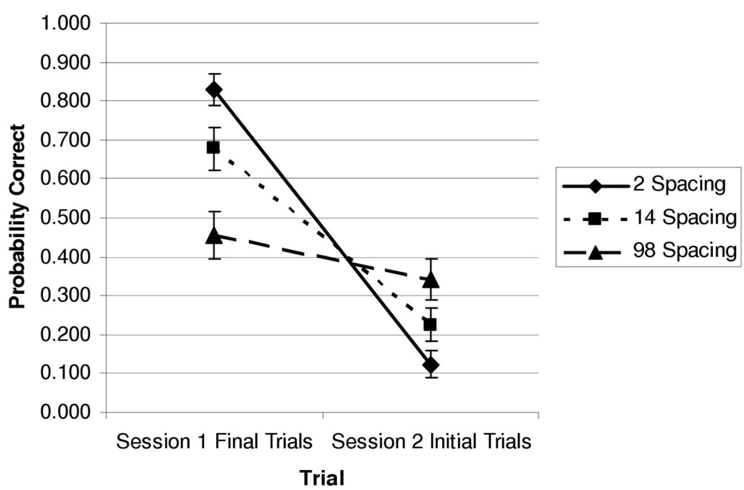


Fig. 2. Spacing Crossover Interaction. S2 initial trial performance and S1 final trial performance as a function of spacing. Values exclude the 8 test trial condition and aggregate all other repetition and retention conditions. 2 *SE* confidence intervals computed from participant means.

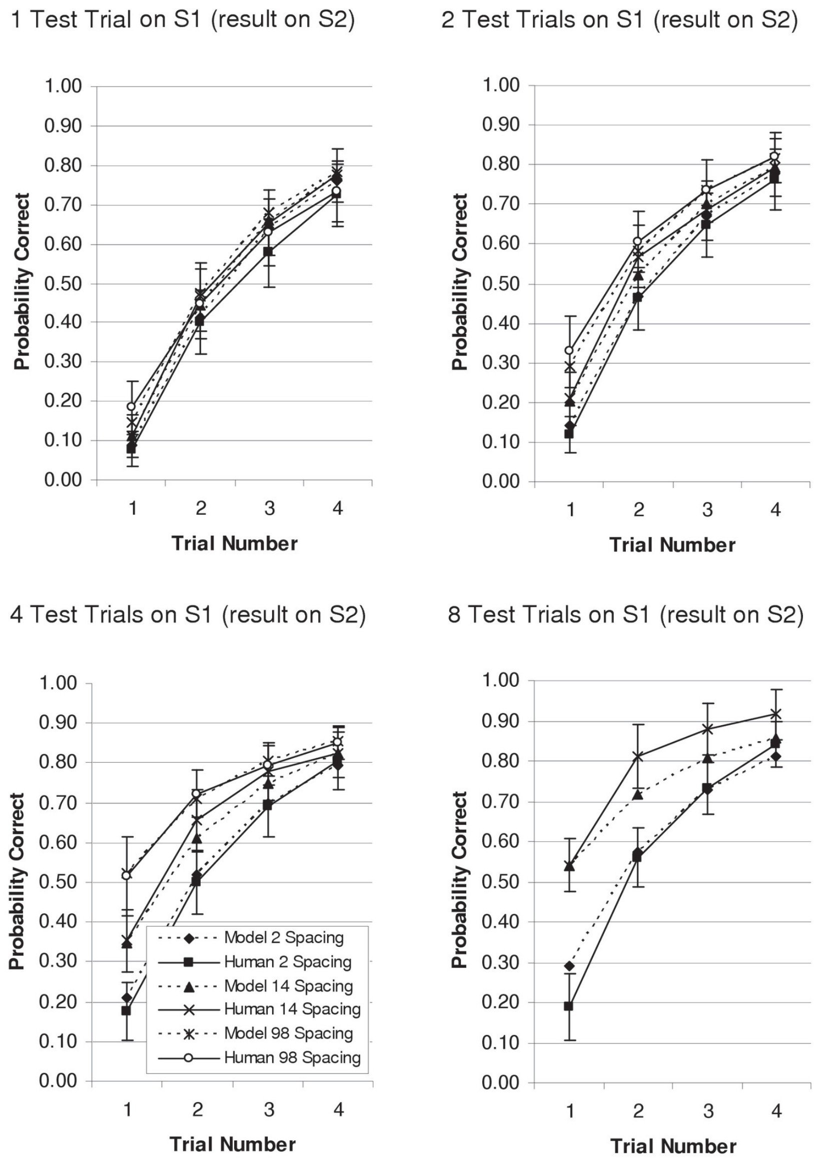


Fig. 3. Experiment 1 S2 aggregate data for humans and model for practice conditions by spacing intervals. Individual graphs for each repetition condition. 2 *SE* confidence intervals computed from participant means.

We were then interested in showing the effects present in the S2 data. The mean correctness for S2 was 0.64 for the 1-day retention interval and 0.52 for the 7-day retention interval. Because the patterns of data were similar between the two retention intervals on S2 ($r = .959$, $p < .001$), we aggregated the two conditions for purposes of display (see Fig. 3). Some main effects and interactions can be noted. A repeated measures ANOVA of S2 data (S1 repetitions × S1 spacing × S2 trial × Retention interval, excluding the eight repetition conditions due to the in-complete design) was completed. First, this analysis showed that people forgot more when the retention interval went from 1 to 7 days, $F(1, 38) = 4.26$, $p < .05$. It also revealed a strong main effect of spacing, $F(2, 76) = 58.2$, $p < .001$. More important, we found a significant S1 repetition × S1 spacing interaction, $F(4, 152) = 4.38$, $p < .005$, reflecting an increasing benefit to spacing with more repetitions at a particular spacing. This is similar to interactions produced by Underwood (1969). As can be seen from Fig. 3, there was a dramatic increase in the importance of spacing as S1 repetitions increased.

## 4. Discussion of model and theory

Our modeling of these data was motivated by the experiment and by results such as Bahrick (1979). Notable in such results are strong crossover interactions. Crossover interactions such as Bahrick's and our interactions (see Fig. 2) suggest that the rate of forgetting is different depending on the spacing of practice. Although statistical analysis of this sort of forgetting-rate interaction are notoriously difficult due to scaling issues (Bogartz, 1990), the crossovers we found would occur regardless of scaling and thus were good evidence that in our experiment forgetting was slower after a series of spaced presentations in comparison to a more massed series.

Another result of interest was the finding that the effects of spacing are greater the more practice trials there are. Combined with the crossover interactions, it implied that the benefits of spacing in slowing the forgetting rate become larger as the number of practice trials increases.

To model these data we needed a formal system that (1) predicted the improvement in performance that occurs with practice, (2) predicted the decrease in performance with delay, (3) predicted the spacing effect, (4) predicted the interaction of spacing with retention (that spaced practice shows greater advantage with greater delay), and (5) predicted the interaction of spacing with practice (that spacing is more important when there are more practice trials). The standard ACT–R model seems well suited to handle the effects of practice and the effects of retention interval. It not only predicts the effects of practice but also the form of the practice and retention functions—that they are both roughly power functions. We review how the theory predicted these effects first, before going on to describe the elaboration that we produced to include the spacing effect and its interactions with practice and retention.

ACT–R's activation equation represents the strength of a memory item as the sum of a number of individual memory strengthenings, each corresponding to a past practice event. It proposes that each time an item is practiced the activation of the item receives an increment in strength that decays away as a power function of time. These individual strengthenings[3] resulted in the following equation for strength of an item after $n$ presentations:

$$m_n(t_{1 \ldots n}) = ln\left(\sum_{i=1}^{n} t_i^{-d}\right) \tag{1}$$

In this function, $m$ is the activation of the item as a function of the times ($t_i s$) since each of the $n$ prior presentations.[4] Each $t_i$ is how long ago the $i$th practice of that item occurred, and these values are scaled to account for differences in interference (this scaling is detailed later). The decay parameter ($d$) is a constant. Combined with the response functions in ACT–R, this activation equation produces the power laws of practice and forgetting. One can note regarding Wickens (1999) that this equation provides a number of the required components of practice and forgetting functions. It defines initial learning given prior practice, and it explains both the progression of forgetting and the integration of the effect of each discrete practice event.

Because ACT–R is a complex system, for this article we model simply the effect of practice and forgetting on a trace that encodes the two items in a paired associate. Further, our model abstracts over issues of cuing and context. In the full ACT–R model, in addition to any effects of practice and forgetting on the memory of the trace encoding the paired associate, spreading activation from the number of cues or from the context affects the activation of a memory. However, because the effect of number of cues and context is constant across trials in our experimental task (because there is only one cue and our context does not fluctuate with time), the predictions we make are equivalent to those from the full activation equation. Given this experiment, if we were to consider the spreading activation term, we would simply need to reestimate $\tau$ (in the following equation) to compensate for the constant increase due to activation spread.

The response function of interest in this experiment concerned accuracy. In ACT–R, an item will be retrieved if its activation is above a threshold. Because activation is noisy, an item with activation $m$ as given by Equation 1 has only a certain probability of recall. ACT–R assumes a logistic distribution of activation noise, in which case the probability of recall is:

$$p_r(m) = \frac{1}{1 + e^{\frac{\tau - m}{s}}} \tag{2}$$

In this equation, $\tau$ is the threshold parameter and $s$ is the measure of noise. An inspection of the formula shows that, as $m$ tends higher, the probability of recall ($p_r$) approaches 1, whereas, as $\tau$ tends higher, the probability decreases. In fact, whev $\tau = m$, the probability of recall is .5. The $s$ parameter controls the noise in activation, and it describes the sensitivity of recall to changes in activation. If $s$ is close to 0, the transition from near 0% recall to near 100% will be abrupt, whereas when $s$ is larger, the transition will be a slower sigmoidal curve. Because this function results in diminishing marginal returns for practice and diminishing marginal losses for forgetting, it address the need for explaining asymptotic performance described by Wickens (1999).

Although the recall probability function explains one aspect of asymptotic forgetting, the slowdown of forgetting over long delays between practice sessions, exemplified by our experiment, is handled by a scaling of the $t_i$ values in Equation 1. J. R. Anderson et al. (1999) found that although the activation equation could account for practice and forgetting effects within an experiment, it was not able to fit retention data over long intervals between sessions (they looked at retention intervals of up to 1 year). Therefore, they found it necessary to suppose that between sessions there is less destructive interference from intervening memory events than during an experimental session. They modeled the apparent slowing of decay by scaling the passage of time outside the experiment. Forgetting is then dependent on this “psychological time" between presentations rather than the real time. This is implemented by multiplying the portion of time that occurred between sessions by the $h$ parameter.[5]

In this theory, then, interference interacts with the decay rate to control the true rate of forgetting. For the models in this article, the scale factor to convert real time to psychological time (a measure of intervening interfering events) is 1 within an experiment and 0.025 between experiments. Therefore, in our memory equations, each $t_i$ represents the cumulative interference a presentation has encountered. Because of this the rate of memory loss for any $t_i^{-d}$ at any time (the hazard rate) is best thought of as being a function of the decay rate and the total amount of interference encountered. Thus, we are supposing that Jost's law applies to interference and might be restated as follows: Given two memories, both of which are currently equal in recall strength, the one that has already suffered the most from interference will be forgotten more slowly.

This intuition from the model allows the theory to provide a bridge between understanding forgetting in terms of time and understanding forgetting in terms of interference. Forgetting in this theory is linked to time, but it also depends on the rate of interference for an interval. Because we have found that $h$ (the interference rate) tends to be stable, we have used the same ratio 1/40 (interference being 40 times greater within an experiment) across the two experiments we fit that have between-session intervals. We would have been able to fit the data for these experiments more tightly if we had estimated different $h$ values for each experiment, or if we had not assumed within-experiment $h$ to be 1 (as we do for all our fits), but conceptually it would have been more difficult to interpret differences in forgetting due to spacing. Because we are focusing on spacing effects more than interference processes here, we decided to forgo investigation of how $h$ might vary for each experiment.

### 4.1. Decay rate as a function of activation

Even with the elaboration offered by J. R. Anderson et al. (1999), ACT–R (J. R. Anderson & Lebiere, 1998) was incapable of predicting any crossover spacing effect. However, J. R. Anderson and Schooler (1991) proposed a modification to the ACT–R strength equation that does capture the spacing effect. This modification specified that each presentation had an individual decay rate that depended on the spacing from the prior presentation. It was a formalization of a mechanism suggested by Wickelgren (1973). J. R. Anderson and Schooler's specific proposal was that the $i$th presentation would have the decay rate:

$$d_i(t_i, t_{i-1}) = \max[d, b(t_i - t_{i-1})^{-d}] \tag{3}$$

In Equation 3, $d$ is the minimum decay rate (which applies for the first presentation and for presentations at long enough lags.) At shorter lags the decay rate for a presentation is itself a power function of the lag, $t_i - t_{i-1}$. Although J. R. Anderson and Schooler (1991) had some success with this form of the equation they commented that "its exact form is a bit arbitrary" and that "there is not evidence one way or the other for this precise" formulation (p. 407). We had at least three other problems. First, it made the decay rate for a presentation just a function of the lag since the last item, and this did not seem to be plausible. Second, our new formulation provided marginally better fits across models. Third, we have found the new formulation has more parameter stability across models. This final point suggests it may better describe the underlying processes.

As an alternative to the J. R. Anderson and Schooler (1991) proposal, we developed an equation in which decay for the *i*th presentation (considering the initial study as Presentation 1), $d_i$, is a function of the activation at the time it occurs instead of at the lag (see Equation 4.) This implies that higher activation at the time of a trial will result in the gains from that trial decaying more quickly. On the other hand, if activation is low, decay will proceed more slowly. Specifically, we propose Equation 4 to specify how the decay rate, $d_i$, is calculated for the *i*th presentation of an item as a function of the activation $m_{i-1}$ at the time the presentation occurred. Equation 5 then shows how the activation $m_n$ after $n$ presentations depends on the decay rates, $d_i s$, for the past trials.

$$d_i(m_{i-1}) = ce^{m_{i-1}} + a \tag{4}$$

$$m_n(t_{1 \ldots n}) = ln\left(\sum_{i=1}^{n} t_i^{-d_i}\right) \tag{5}$$

In Equation 4, $c$ is the decay scale parameter, and $a$ is the intercept of the decay function.[6] For the first practice of any sequence, $d_1 = a$ because $m_0$ is equal to negative infinity. Note that when $c = 0$, Equation 4 is nullified, and Equation 5 collapses to the standard ACT–R Equation 1. These equations are recursive because to calculate any particular $m_n$ one must have previously calculated all prior $m_n s$ to calculate the $d_i s$ needed. The Appendix includes a detailed example applying these equations for a sequence of five presentations. These equations result in a steady decrease in the long-run retention benefit for additional presentations in a sequence of closely spaced presentations. As spacing gets wider in such a sequence, activation has time to decrease between presentations; decay is then lower for new presentations, and long-run effects do not decrease as much.

The original ACT–R model (Equation 1) does not produce the spacing effect because it has no mechanism to reflect that time differences between practices should matter much. The spacing effect in this model (Equations 4 and 5) occurs because when spacing between two presentations is wider, the decay rate for the second presentation is lower. At long retention delays, this more than compensates for the fact the first presentation suffers more forgetting due to the increased retention interval from the wider spacing.

The new model shows that each practice contributes to a single unitary strength measure for the represented chunk. However, the activation equation captures this overall strength as a number of discrete contributions represented explicitly by the contribution from each $t_i^{-d_i}$. A neural analogy for this relation might suppose that each experience ($t_i$) results in the creation of new receptor sites at the synapses that correspond to the overall memory trace. Indeed, an addition of synapses with the induction of long-term potentiation (LTP) of neural connections has been shown experimentally. For instance, Toni et al. (2001) and Geinisman (2000) showed that LTP induction results in perforated areas on dendritic spines, which later develop into multisynapse connections with presynaptic cells. Presynaptic cells undergo a similar activity-dependent remodeling that results in new axonal synapses (Nikonenko, Jourdain, & Muller, 2003).

The model then says that the stability of these new receptor sites (we can consider the $-d_i$ value as a measure of stability) is less when they are created when strength is already high. In general, this argument supposes that the faster forgetting following massed practice may reflect diminishing marginal returns in the initiation of the neural consolidation processes.[7] Indeed, one can see that decay and consolidation are analogous in the theory because the $d_i$ parameter can characterize either depending on its sign. Like decay, consolidation can be considered a continuous process that is initiated at encoding and involves memories becoming more stable over time (in conformance with Jost's law), but they are simultaneously degrading due to forgetting processes.

We attempted to fit a model using Equations 2, 4, and 5 to the data from the experiment. All fits for this article were performed by minimizing a $\chi^2$ statistic computed from the condition means according to the formula

$$\chi^2 = \sum_i \frac{N_i(pred_i - obs_i)^2}{pred_i - pred_i^2} \tag{6}$$

Where the summation is over the $i$ data points, $N_i$ is the number of observations for each data point, $pred_i$ is the predicted recall probability in condition $i$, and $obs_i$ is the observed probability. We should note that this statistic does not satisfy the assumptions of the chi-square distribution because of nonindependence of observations. However, minimizing it is still a reasonable way to estimate parameters, and its use allows comparison with Raaijmakers (2003), who used this statistic in the same way. As an alternative to evaluate and compare fits we have also provided $r^2$ and root mean square deviation (RMSD) statistics. In this article, the RMSD values were adjusted for model complexity by subtracting the number of parameters from the divisor when computing the mean, as described in Pitt, Myung, and Zhang (2002).

We applied this extended ACT–R model to the experiment to fix parameters, which we tried to preserve in fitting other data sets. For this experiment, there were 162 aggregate correctness averages (the $obs_i$s in the previously mentioned chi-squared summation) to be fitted and split into 81 points for each between-subject retention interval, of which 37 points were for the S1 conditions, and 44 points were for the S2 conditions. Table 1 gives the parameters and goodness-of-fit measures for this model. As is apparent from Figs. 1 and 3, the model mirrored the absolute and relative patterns in the data closely.

Applying the model to the experiment allowed us to fix default parameters for the model. Using these defaults, we then adopted the policy of keeping as many parameters as constant as possible between models. Because this policy limits the complexity of our model, we believe it improves its explanatory utility.

### *4.2. ACT–R versus contextual fluctuation (SAM)*

Raaijmakers (2003) extended the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1981) to account for the spacing effect and successfully fit the model to some data sets. Therefore, we decided to compare Raaijmakers' model with ours using the data from our experiment and three experiments that Raaijmakers reported fits to. His model has two mechanisms that cause spacing effects. First, it has a short-term store (STS) mechanism that contains recently encoded information. If an item is still in this STS at the time of a later presentation, the new presentation does not get a second encoding. This creates a spacing effect because at longer

Table 1
Parameters and statistics for all data sets

| | Parameters and Model Statistics | | | | | |
|---|---|---|---|---|---|---|
| | Experiment | Experiment Reduced Model | Bahrick (1979); Bahrick and Phelps (1987) | Rumelhart (1967) | Young (1971) | Glenberg (1976) |
| Parameters | | | | | | |
| Decay intercept ($a$) | 0.177 | 0.172 | 0.217 | 0.149 | 0.300 | 0.058 |
| Decay scale ($c$) | 0.217 | 0.250 | 0.143 | 0.495 | 0.419 | 0.283 |
| Threshold ($\tau$) | –0.704 | –0.704[a] | –0.704[a] | –0.704[a] | –0.704[a] | –0.704[a] |
| Noise ($s$) | 0.255 | 0.255[a] | 0.255[a] | 0.255[a] | 0.255[a] | 0.255[a] |
| Inteference scalar ($h$) | 0.025 | 0.025[a] | 0.025[a] | n/a | n/a | n/a |
| Encoding scalar (b) | 1[a] | 1[a] | 3.79 | 1[a] | 1[a] | 0.352 |
| Reduced encoding ($b_r$) | n/a | n/a | n/a | n/a | n/a | 0.274 |
| Fit Statistics | | | | | | |
| $r^2$ | 0.944 | 0.990 | 0.926 | 0.927 | 0.461 | 0.944 |
| RMSD adjusted | 0.046 | 0.031 | 0.060 | 0.021 | 0.026 | 0.026 |
| $\chi^2$ | 328 | 43.3 | 258 | 41.5 | 8.70 | 31.9 |
| $\chi^2$ *df* | 157 | 28 | 27 | 38 | 16 | 20 |

*Note.* RMSD = root mean square deviation.
[a]Fixed parameters.

spacings an item is more likely to have left STS and to have been encoded. The second mechanism involves contextual fluctuation such that the contextual elements available for incorporation into memory traces fluctuate over time. This creates a spacing effect because when presentations are more widely spaced the resulting trace will be composed of a more varied sample of the possible contextual elements. In the SAM model then the probability of recall is a function of the overlap between this memory trace and current test context. Both of the mechanisms the model uses, failure to reencode at short lags and contextual overlap at shorter spacings, have been ideas that have been mentioned in a number of proposals about the spacing effect (e.g., Atkinson & Shiffrin, 1968; Glenberg, 1979). Therefore, Raaijmakers' model represents an attempt to formalize these ideas and offers us a measure of how well they can account for our results.

Ross and Landauer (1978) pointed out that the contextual fluctuation theory, because it proposes samples of the environment that are independent, predicts that a spacing effect should occur for the recall of either or both of two different items presented at a spacing, just as it does for the same item presented twice at a spacing. They performed three experiments that convincingly showed that this is not the case. Glenberg and Smith (1981) agreed that this was a major problem for the contextual fluctuation theory. However, Raaijmakers's (2003) model introduced a crucial change to the theory of contextual fluctuation by supposing that "study-phase retrieval" was necessary for the information from a new presentation to be added to the trace for that item. This solves the problem of Ross and Landauer (1978) because the second of two different items cannot trigger retrieval of the first; however, it might be noted that it does so by introducing a dependency between presentations (because trace cumulation requires recall) that makes the model seem somewhat different from the independent encoding proposition Estes (1955) and Glenberg (1979) proposed.

To enable this model to handle long-term memory data from our experiment we added a mechanism like the "psychological time" mechanism in our model. This allowed the model a free parameter to find the best fitting number of trials to represent the forgetting over the time between sessions. With this adjustment it was possible to take the model that Raaijmakers (2003) developed for a model of Rumelhart's (1967) data, adjust it to reflect the presentation schedule in our experiment, estimate new parameters, and make predictions. However, we encountered a couple of difficulties in fitting the Raaijmakers model to our data.

The first issue was purely technical. Because of the combinatorial complexity of the Raaijmakers model, it was impractical to model the eight-test trial condition. Therefore, we excluded these data from our modeling effort. To further limit the combinatorial complexity, we only looked at first trial performance on S2. The second issue was more complex. The basic Raaijmakers model was simply unable to fit the S1 learning data in the 98-spacing condition. This is largely because retrieval during practice is so poor with 98 spacing that the trace cumulation mechanism fails to result in the memory gains observed. Because including this condition in the Raaijmakers model seriously distorted the parameter estimation, and still resulted in poor fits, we altered the model slightly at Raaijmakers suggestion (personal communication, March 8, 2004), so the study-phase retrieval mechanism occurred automatically. This made it so that the repetitions were always recognized and trace cumulation could not fail.

With this modification the Raaijmakers (2003) model was able to cope with the 98-spacing condition and produce a reasonable overall fit ($\chi^2 = 166$, $df = 23$). However, the ACT–R model resulted in a better fit ($\chi^2 = 43.3$, $df = 28$) with the same restricted data set (see Tables 1 and 2). Both models did fairly well in capturing the main effects and interactions in the data. The problem for the Raaijmakers model fit is still with the nature of the growth function. Fig. 4 tries to

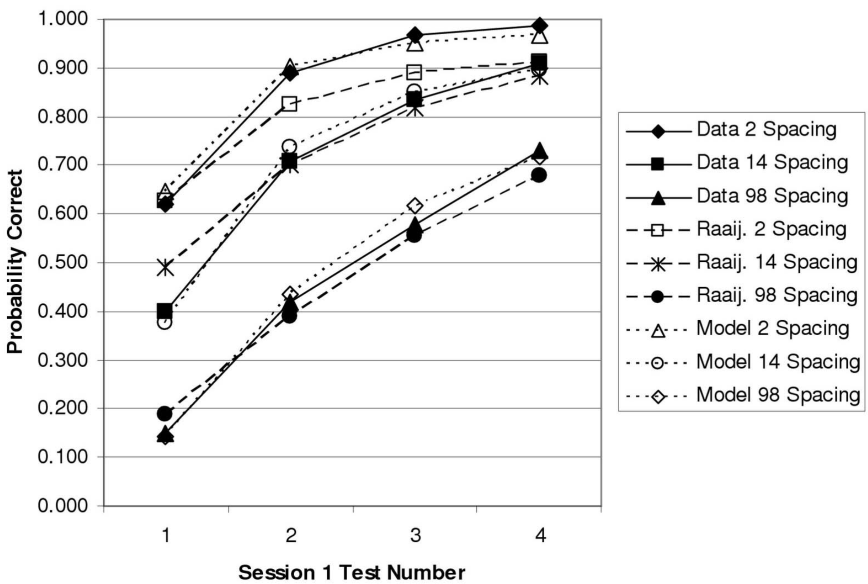


Fig. 4. Comparison of data and fits by Raaijmakers model and ACT–R model.

show what is behind the lack of fit of the Raaijmakers model. (The complete Raaijmakers model fit is available at the Web site.[1]) In Fig. 4, we have averaged over the 1, 2, and 4 practice conditions to look at the forms of the learning curves. One can note that the Raaijmaker's model has a problem with capturing sufficient Final Session 1 learning across all the conditions. In part, particularly for the two-spacing condition, this may be caused by the STS mechanism blocking encoding when an item remains in STS on repetition. Because of this mechanism, closely spaced practices do not gain much contextual strength because of overlap and because they are sometimes not encoded. It appears that the model cannot compensate for this by adjusting the contextual overlap parameters, likely because this would upset the fit of the 98-spacing condition in which the STS mechanism plays no role. To check the possibility that this problem was due to issues involving the fit to Session 2 tests, we also ran the model for only Session 1 data; this did not greatly improve the fit to the learning curves ($\chi^2 = 104$, $df = 14$).

Later in this article, we present comparisons of our model fits with the Raaijmakers (2003) model fits for a set of three experiments in the literature. Our model fits these data sets comparably well and with fewer parameters. It seems reasonable to infer that the Raaijmakers model and the ideas on which it is based can predict many of the basic effects of spacing, as our model can. However, the following experiments did not measure retrieval performance over as large a range of correctness values as our experiment did, and therefore this issue with the learning function was not noticed. Thus, our current experiment brought out a critical advantage of our model in capturing the learning curves during spaced practice.

## 5. Other model fits

To test further the generalizability of our model, we fit it to four examples from the memory literature. For each of these fits we varied as few parameters as possible, preferring instead to use the defaults from the experiment. The first experiment involved long-term retention intervals, whereas the following three experiments involved only a single session. These three single-session experiments were also modeled by Raaijmakers (2003), and we briefly compare our fits of these experiments with his fits.

### *5.1. Bahrick*

Bahrick's (Bahrick, 1979; Bahrick & Phelps, 1987) results on learning Spanish–English vocabulary pairs are seminal in discussions of the spacing effect, and therefore we wanted to show that our model could account for them. In Bahrick, participants memorized 50 Spanish–English vocabulary pairs. The training took place over three or six sessions. Each session began with testing of all the words followed by a presentation sequence of any word pairs not recalled. After presentation, these words were retested and the procedure repeated until all words were correctly recalled once in each session. The experimental sessions were spaced every 0, 1, or 30 days and were followed by a final test session at a 30-day retention interval. Table 2 shows the performance during the initial testing sequence of each session. The first training session did not begin with testing, and thus it is not listed. We have also included the data

Table 2
Bahrick (1979) and Bahrick and Phelps (1987) human data and model fits

| | Session | | | | | | |
|---|---|---|---|---|---|---|---|
| Spacing (days) | 2 | 3 | 4 | 5 | 6 | 30 days | 8 years |
| Human results | | | | | | | |
| 0 | 0.77 | 0.89 | | | | 0.33 | |
| 1 | 0.60 | 0.87 | | | | 0.64 | |
| 30 | 0.21 | 0.51 | | | | 0.72 | |
| 0 | 0.82 | 0.92 | 0.96 | 0.96 | 0.98 | 0.68 | 0.06 |
| 1 | 0.53 | 0.86 | 0.94 | 0.96 | 0.98 | 0.86 | 0.08 |
| 30 | 0.21 | 0.51 | 0.72 | 0.79 | 0.82 | 0.95 | 0.15 |
| Model results | | | | | | | |
| 0 | 0.80 | 0.92 | | | | 0.55 | |
| 1 | 0.62 | 0.84 | | | | 0.62 | |
| 30 | 0.19 | 0.49 | | | | 0.66 | |
| 0 | 0.80 | 0.92 | 0.95 | 0.97 | 0.97 | 0.77 | 0.05 |
| 1 | 0.62 | 0.84 | 0.90 | 0.93 | 0.95 | 0.83 | 0.07 |
| 30 | 0.19 | 0.49 | 0.66 | 0.76 | 0.81 | 0.85 | 0.18 |

from Bahrick and Phelps. These data involved recalling available participants from the 1979 study for retest on the same material.

Our basic task was similar to the Bahrick (1979) task, and we also had fairly long spacings between sessions. However, the training was different. Bahrick used a successive dropout strategy that resulted in a variable number of practices with each word pair for a particular session. Further, participants had both visual and verbal presentations of words in the Bahrick study. To account for this stronger encoding within each session, for our model we assumed a single practice with each word for each session but estimated a strength multiplier of 3.79, which accounted for the extra time for encoding and the possibility of multiple encodings within a session. This multiplier (the $b$ parameter) was used to scale the individual $t_i^{-d_i}$ s in the activation function so each $t_i^{-d_i}$ contributed 3.79 times what it would normally. Further, for the 0 spacing between sessions condition we estimated that 40 min elapsed from the beginning of one session until the beginning of the next. This was not a fitted parameter but was a reasonable guess based on Bahrick's methods. We fitted the decay parameters ($a = 0.217$ and $c = 0.143$) and encoding strength parameter, $b$, by minimizing the $\chi^2$ statistic ($\chi^2 = 258$, $df = 27$).

Table 2 shows that the model's results were similar to the participants' data. More important, the crossover interaction across the final 30-day retention period was captured by the model. The $r^2$ was .926 with an adjusted RMSD of 0.060. One problem in fitting came during the final 30-day test, in the six-session data, where the model underestimated how well participants did in the 30-day lag-retention condition. However, it can be noted that participants in this condition showed a jump in their recall in the final test (i.e., participants increased 3% from Session 5 to 6, and then jumped up 13% to the final 30-day test), suggesting that this data point may be somewhat anomalous. It can also be noted that our model failed to predict well during the final 30-day test when there had been three sessions with zero lag. On the other hand, the good fit of the 8-year retention data provided evidence that the long-term forgetting mechanism (using the *h* parameter) behaves consistently even at very long intervals.

### 5.2. Rumelhart (Experiment 1)

Rumelhart (1967; Experiment 1) is an example of an experiment that tested both our combination rule for multiple practices and our mechanism for spacing. It is also a data set fit by Raaijmakers (2003). In the experiment, participants performed a continuous paired-associate recall task with 66 different items including fillers. Eight different sequences of spacing were used, and each sequence was used six times across the experiment. The stimuli consisted of consonant–vowel–consonant trigrams paired with a digit, either 3, 5, or 7. Each trial consisted of a test with the stimulus, a 2-sec presentation of the stimulus–response pair, and a 3-sec intertrial interval.

Fig. 5 presents the data Rumelhart (1967) collected and our model of the data. The participants' results are similar to our experiment. For example, in the 10–10–10–10–10 condition participants learned relatively slowly compared with the 1–1 – 1–1 – 10 condition. However, as our model predicts, after four trials 1–1 – 1–1, the forgetting until the final test after 10 trials was pronounced compared with the forgetting for the final test after 10 trials with 10–10–10–10 spacing. This is evidence that wider spacing resulted in more stable learning as our model implies. First trials are not listed in Fig. 5 because these responses were at chance levels because participants had no prior practice.

To model these data we assumed each trial was 10 sec long based on Rumelhart's (1967) methods assuming 5 sec per response. Because guessing had a one-third chance of success, this was included in the model. We estimated the decay parameters ($a = 0.149$ and $c = 0.495$) to minimize the $\chi^2$ ($\chi^2 = 41.5$, $df = 38$). These decay parameters indicated a relatively high forgetting rate. This was reasonable considering the stimuli were arguably unmemorable consonant–vowel–consonant nonwords, and the responses were easily confusable.

This fit captured the important effects in Fig. 5 (adjusted RMSD = 0.021, $r^2 = .927$). Because our model had been developed with different stimuli over durations of days, we considered the good fit to this data with only two free parameters estimated as evidence in support of our model. These data have also been modeled by Raaijmakers (2003; See Table 3). The fit was good ($\chi^2 = 38$, $df = 34$), but note that six parameters were varied. Given the degrees of freedom difference in the models, this fit was roughly equivalent to the ACT–R fit.

### 5.3. Young (1971)

In all of our examples so far, wider spacing resulted in better performance later. Young (1971) was one of the first to provide a demonstration that spacing does not always affect performance monotonically. In a continuous paired-associate memory experiment, he paired consonant trigrams with single digits. Each pair was presented twice for study, with a spacing interval of 0 to 17 trials (study trials were 1 sec in duration with a 3-sec intertrial interval). The retention interval was held constant at 10 trials (40 sec) from the second practice.

The data can be seen in Fig. 6 with a comparison to a fit by the model. This fit captured the nonmonotonic nature of the spacing effect. This result makes sense, because if the spacing is

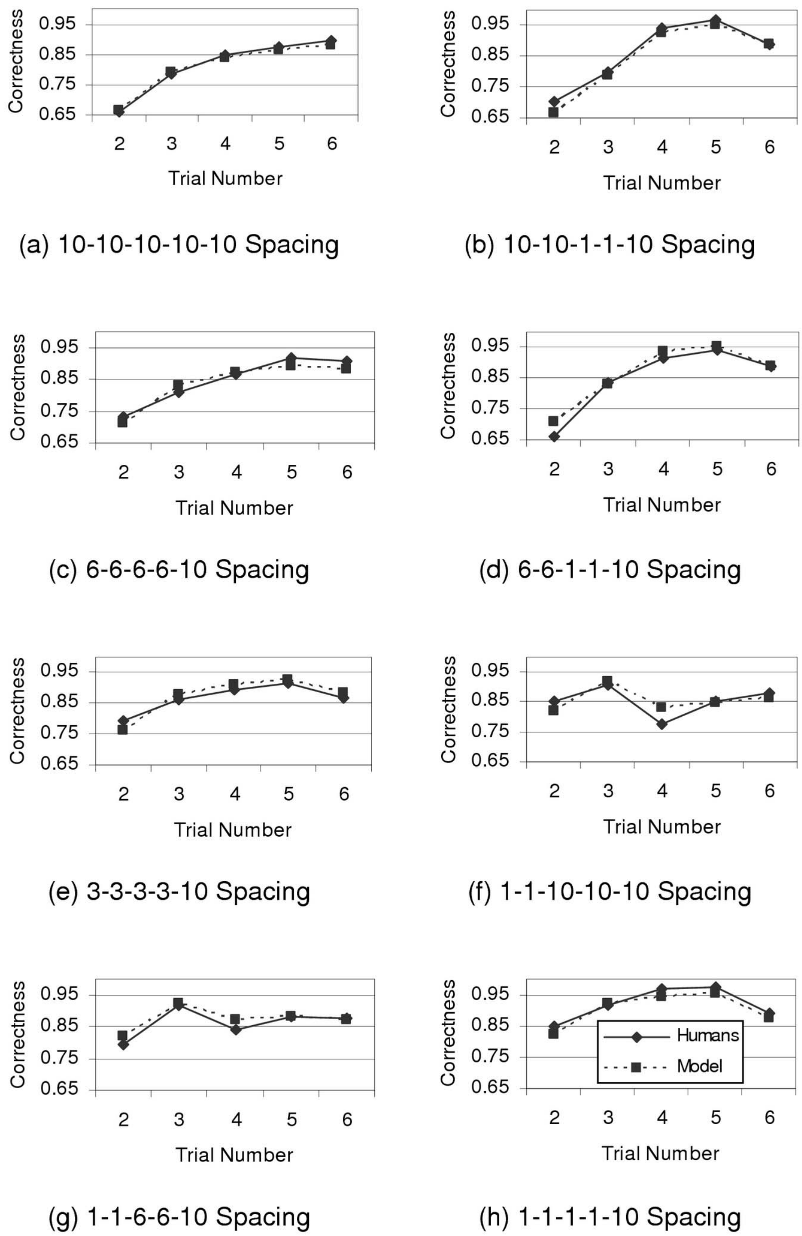


Fig. 5. Human and model data for Rumelhart (1967).

too short, the second presentation will be forgotten rapidly, whereas if the spacing is too long, the strength contribution of the first trial will have declined due to the longer retention interval. This implies a specific ideal spacing interval given a particular retention interval. For the Young (1971) data it appeared that the best retention occurred when the spacing was roughly six trials. Given the best fitting parameters ($a$ = 0.300 and $c$ = 0.419), our model agreed with

Table 3
Raaijmakers model parameters and fit statistics

| | Parameters and Model Statistics | | | |
|---|---|---|---|---|
| | Experiment Reduced | Rumelhart (1967) | Young (1971) | Glenberg (1976) |
| Parameters | | | | |
| Fluctuation parameter ($a$) | 0.013 | 0.087 | 0.082 | 0.013 |
| Fluctuation parameter ($s$) | 0.047 | 0.288 | 0.150 | 0.260 |
| Scaling constant for context association ($a$) | 5[a] | 5[a] | 5[a] | 5[a] |
| Probability that an item enters the STS buffer ($w$) | .666 | .766 | 1[a] | 1[a] |
| Interitem information stored on a first study trial ($b$) | 0.430 | 0.688 | 0.246 | 0.732 |
| Interitem information stored on subsequent trials ($b_2$) | 0.570 | 0.688 | 0.246 | 0.732 |
| Interference constant ($Z$) | 3.0 | 3.0 | 2.0 | 10.0 |
| Scaling parameter in recovery equation for a test trial ($\theta_2$) | .5[a] | .5[a] | 0.3 | 0.215 |
| Rate of decay from STS ($\lambda$) | 0.128 | 0.310 | 0.746 | 0.800 |
| Retrieval attempts ($L_{max}$) | 3[a] | 3[a] | 3[a] | 3[a] |
| Trials estimated between sessions | 240 | | | |
| Fit Statistics | | | | |
| $\chi^2$ | 166 | 38.0 | 8.47 | 41.9 |
| $\chi^2$ *df* | 23 | 34 | 12 | 18 |

*Note.* STS = short-term store.
[a]Fixed parameters

this figure ($\chi^2 = 8.70$, $df = 16$). These parameter values imply rapid forgetting, which is plausible for consonant trigrams. The RMSD was 0.026. The low $r^2$ of .461 reflected the noise in Young's data relative to the basic curvilinear trend.

Raaijmakers (2003) was also successful in fitting this data ($\chi^2 = 8.47$, $df = 12$). However, again, six parameters were varied.

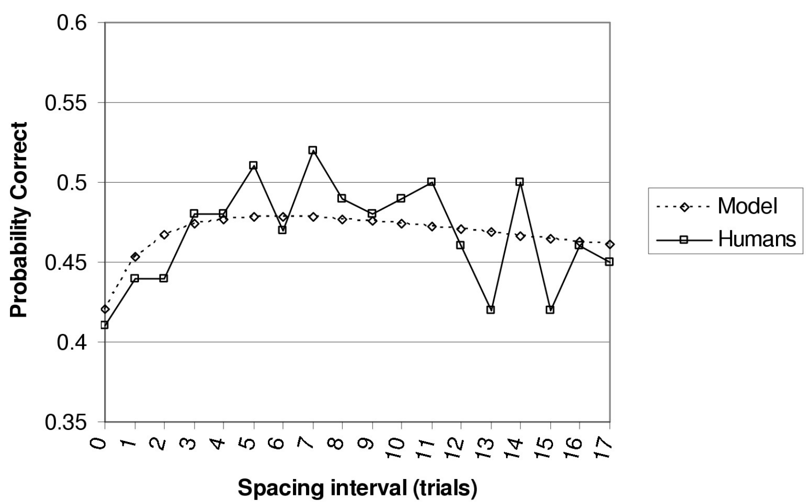


Fig. 6. Young (1971) human data and model fit.

### 5.4. Glenberg (1976; Experiment 1)

Glenberg (1976; Experiment 1) has been modeled by many individuals interested in the spacing effect (J. R. Anderson & Schooler, 1991; Raaijmakers, 2003; Reed, 1977). This recurrent interest is due to the interesting monotonic and nonmonotonic effects across the four retention intervals and six levels of spacing examined. In the experiment, participants were presented with pairs of unrelated common nouns. Each pair was presented, without testing, twice at lags of 0, 1, 4, 8, 20, and 40 trials. Following these two presentations, after a 2-, 8-, 32-, or 64-trial retention interval, there was a test trial during which participants were cued with the first word of the pair and responded with the second. Presentations and tests were 3 sec long. Accuracy on these tests across the conditions is plotted in Fig. 7.

Notable in the data is the reversal from a nonmonotonic result to a monotonic spacing effect as the retention interval went from short to long. Thus, we have a single experiment showing both the nonmonotonic and monotonic spacing effects reported in the literature. Similar to the explanation in the Young (1971) section, we propose that this effect is caused by slowed forgetting of the second presentation at longer lags. This slowed forgetting does not have time to result in much of an effect at short retention intervals, so performance is best when the spacing is narrower. On the other hand, at longer retention intervals, the differences in forgetting have an increasing influence, and the ideal spacing appears to be at least 40 trials when the retention interval is 64 trials. This defines an important effect of retention interval on optimal spacing, because the optimal spacing increases monotonically with the increase in retention interval.

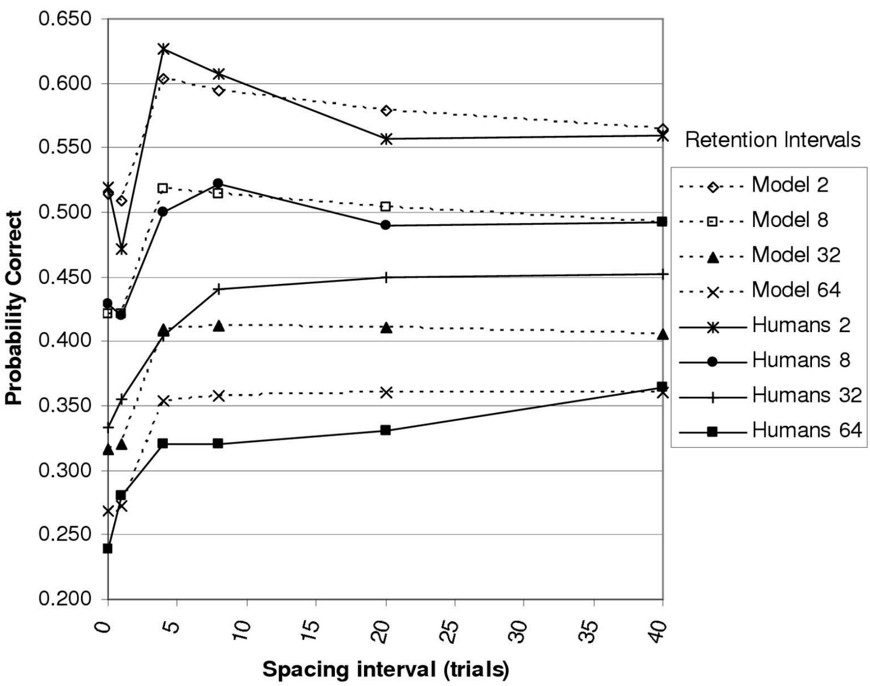


Fig. 7. Glenberg (1976) human data and model fit.

In capturing this experiment with our model, we needed to deal with two issues in the design of Glenberg's (1976) Experiment 1. Both of these issues involved the speed of presentation in the experiment. First, in our experiment, which also used verbal stimuli, the presentation–test interval (including intertrial time) was approximately 5 to 10 sec, whereas for this experiment it was 3 sec long. Because of these shorter presentations, we scaled the contribution of each presentation by the *b* parameter, in a fashion identical to the Bahrick (1979) model. Second, we found that the low performance at lags of 0 and 1 trials (see Fig. 7) was not fitted well by our model. We thought this problem might have been caused by poor encoding of second presentations due to first presentations still being in working memory at these lags of only 0 or 3 sec. According to ACT–R, this item leaves working memory (the goal buffer) automatically on the beginning of each new trial at which point it is counted as an encoding. Further, our model normally assumes that each encoding counts equally at its inception, and only decay causes differences in the long-run effects of presentations. Normally these assumptions cause no problems, given a reasonable lag. However, at the very short lags in this experiment, given the relatively easy Glenberg (1976) material (unlike Young's experiment, which had short spacings but much more abstract material), we needed to model the possibility that items did not drop from working memory as fast as we assumed and thus blocked the full effect of the repetition encoding.

To model this possibility of poorer encoding at these very short lags, we used a reduced $b_r$ parameter to scale the contribution of these second presentations. This solution was similar to the STS mechanism in Raaijmakers's (2003) model and was necessary to get a quantitative good fit. Although this mechanism for extremely short lags was ad hoc, we could have parameterized it as an STS mechanism in a way similar to Raaijmakers. We choose not to do this because we have no other firm evidence for this mechanism except at these very short lags in Glenberg's experiment. Indeed, as Raaijmakers notes, Van Winsum-Westra (1990) was unable to replicate these dips in performance at very short spacing.

Given these changes, the model fit the data well ($\chi^2 = 31.9$, $df = 20$). The RMSD was 0.026 and the $r^2$ was .944. In comparison, the Raaijmakers (2003) model resulted in about the same fit ($\chi^2 = 41.9$, $df = 18$); however, it should be noted that when Raaijmakers made an assumption similar to ours (that 0 lag repetition results in no increase in memory strength) to deal with the short spacing dips, his $\chi^2$ value went down to 28.77. Table 1 shows the parameter values, and Fig. 7 shows the fit of our model to the data. Although the result shows some deviation, it is notable that our model captures the effect of retention interval on optimal spacing (and it captured this effect before we made any of the changes in the preceding paragraph). Raaijmakers' (2003) model did not capture this interaction. Further, his model used two more parameters.

## 6. General discussion

Our experiment produced data to test alternative models of practice, forgetting, and the spacing effect. These data confirmed the standard spacing effect in various conditions and showed that wide spacing of practice provides increasing benefit as practice accumulates. Further, the strong crossover interactions produced provided evidence that people forget less when presentations are widely spaced. The findings of this experiment were used to extend ACT–R's activation equation by introducing a variable decay-rate function. According to this mechanism, the forgetting rate for each presentation of a memory chunk is a function of the activation of the chunk at the time of the presentation. Using this model, we fitted data from our experiment and four experiments from the literature. These fits demonstrated the viability of our mechanism and showed that with the other ACT–R equations it provides an accurate model in a wide variety of conditions. To show that our model was at least as good as an alternative model, we compared our fits for some experiments with fits of the Raaijmakers (2003) model. We were able to produce comparable fits to existing experiments, a better fit to our own experiment, and overall our models had less variation in fewer parameters.

The graphs that we showed for our own experiment aggregated over the two retention intervals, so it may not be immediately clear that the crossover spacing by retention-interval interaction occurred rapidly after final test trials. Fig. 8 shows the predictions of our model for recall at various retention intervals for each spacing condition. These model values and the observed recall at 1- and 7-day retention intervals show the performance we predicted or observed on initial test trials for sessions begun at these retention intervals. This figure makes it clear that the crossover interaction has occurred before the 1-day retention interval session begins. The speed with which the crossovers occur makes sense, given our proposal that different power-law decay values control each retention function. Because the loss rate of memory slows down in power-law forgetting, the greatest changes in strength between the conditions should occur soon after learning.

Given the properties of our model, it is interesting to speculate about what physiological mechanisms might be producing these effects. There are some suggestions that neural plasticity has the features of our model. For instance, in regard to our new decay mechanism, there is work that shows that LTP of neurons declines less rapidly when there is spaced induction of LTP rather than massed induction (Scharf et al., 2002). This result is similar to Wu, Deisseroth, & Tsien (2001) in which spaced stimulation resulted in dendritic changes consistent with

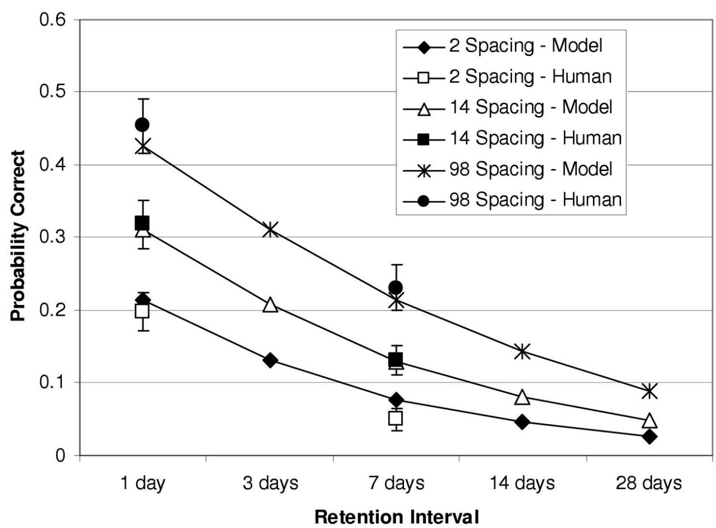


Fig. 8. Predictions of the model (including the relevant data points) for first trials of second sessions begun after the respective retention intervals. Values exclude the eight-test trial condition and aggregate repetition conditions. 1 *SE* confidence intervals computed from participant means.

long-term memory formation, whereas massed stimulation did not have such an effect. Properties of LTP may also correspond to other aspects of our model. For instance, Beggs (2000) presented a statistical model of LTP, proposing that the magnitude of the LTP induced by stimulation is negatively related to this postsynaptic activation. This occurs because increases in LTP in his model are controlled by the discrepancy between presynaptic input and postsynaptic activation. This discrepancy is reduced with LTP induction, and thus subsequent changes in LTP are less. Our model captures this principle by taking the logarithm of the combined trace, thus new encodings add less to activation if it is already high.

In fact, Landauer (1969) has proposed a neural consolidation theory of the spacing effect. In this theory, presenting an item (P1) results in a "hyperexcitable" state in the nervous system following the presentation. This decaying hyperexcitability, according to Landauer, gradually promotes changes in the nervous system responsible for permanent representation of the association. During this period of consolidation, a new presentation (P2) of the same pairing will interrupt the consolidation of P1 due to systemic limits on hyperexcitability. Because of this, the less spacing of presentations, the less memory is strengthened. As Hintzman (1974) pointed out, consolidation theory suffers from a lack of agreement with data that show it is P2 learning that suffers rather than P1 when spacing is narrow. Our version, by placing the effect at P2 rather than P1, no longer suffers from the problems that Hintzman discussed.

Discussing a possible neural basis of the spacing effect suggests an involuntary process. Hintzman (1974) took the broad generality of the spacing effect as evidence that it is not under voluntary control. He proposed that habituation with a stimulus caused it to have less of an effect on increasing long-term memory. Because habituation decreases with time, spaced trials incur a benefit. Although he suggested that massed practice would result in "a decrease in the strength of any new trace that is formed" (Hintzman, 1974, p. 90), he also recognized that the effect occurred for the storage of long-term memories. Thus, although he did not give a formal model, his idea that habituation controls the long-term strength benefit of a spaced practice is similar to our proposal that activation controls the forgetting rate.

The model we have described also agrees with cognitive theories of the spacing effect, which say that the benefit of additional practice is mediated by the difficulty or accessibility of that additional practice. This sort of accessibility theory has been advocated by various researchers (Cuddy & Jacoby, 1982; Schmidt & Bjork, 1992; Whitten & Bjork, 1977). These researchers have noted that manipulations that cause slower acquisition often result in better long-term retention. This paradox occurs in many experimental situations. For instance, Schneider, Healy, and Bourne (2002) conjectured that the greater difficulty participants had learning foreign language responses (as opposed to English responses) in a paired-associate experiment may have produced better long-term recollection. This result makes sense if we suppose greater difficulty indicates lower activation. Thus, a manipulation that increases difficulty might be modeled as a penalty to activation. This lower effective activation would result in less forgetting and therefore better long-term retention.

Although this notion that less accessibility at practice results in better long-term recall is not a new idea, we have presented here the first detailed computational model of how this might occur. We feel this model serves to clarify the theoretical discussion about the effects of practice by allowing clear quantitative and qualitative comparisons to other theories of practice and forgetting such as those proposed by advocates of contextual fluctuation mechanisms. This formal model of the relation between current memory accessibility and the stability of new encodings, using an integrative retention function, provided good fits to a wide variety of results with estimation of only a minimal number of parameters. Further, the absolute and relative parameter stability of the model shows that the model's behavior was consistent with different data sets. This parameter stability enhances the explanatory utility of the model for addressing the broad theoretical issues underlying practice and forgetting.

## Notes

1. Exact trial order for the experiment and working models for all experiments can be found at http://act-r.psy.cmu.edu/models/.
2. This is different from the typical paired-associate procedure that usually also includes a study after successful recall. In a recent study (Pavlik & Anderson, 2004), we compared our procedure with the typical procedure and found long-term differences of less than 2% in recall after multiple practices. Because a study after a successful recall is at a very short spacing, this result is consistent with our model.
3. In ACT–R individual strengthenings are considered to be discrete when the encoding interval is reasonably long. This corresponds to data, such as Melton (1970), which show that as the presentation interval increases, the length of the presentation interval has less effect on final performance. Encoding appears to have quickly diminishing marginal returns. We introduce a $b$ parameter that scales the contribution of each $t_i^{-d_i}$ in the fits for Glenberg (1976) and Bahrick (1979) later in this article to deal with specific methods used in these experiments. This $b$ value is equal to 1 in a standard ACT–R model.
4. The logarithm of the sum is taken to yield observed retention functions and provides a correspondence with log odds of items occurring in the environment as shown by J. R. Anderson and Schooler, 1991. For an extensive review of the mathematical characterization of this system, the reader is referred to Chapter 3 of *The Atomic Components of Thought* (J. R. Anderson & Lebiere, 1998, Appendix A).
5. It should be noted that the psychological time factor in this article takes a slightly different form as compared to J. R. Anderson et al. (1999). In this conception, we take the $h$ factor to be a direct scaling parameter of the time between experimental sessions.
6. Note that because retrieval time in ACT–R is proportional to $e^{-m}$, Equation 4 makes decay an inverse function of retrieval time.
7. The idea that neural consolidation processes depend on spacing is supported by several sources such as Scharf et al. (2002) and Wu et al. (2001).

## Acknowledgments

This research was funded by National Science Foundation Grant BCS 997–5-220. Portions of this work were presented at Fifth International Conference of Cognitive Modeling, Bamberg, Germany.

## References

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1120–1136.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory and Cognition, 25,* 724–730.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: II* (pp. 89–195). Oxford, England: Academic.

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

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 344–349.

Beggs, J. M. (2000). A statistical theory of long-term potentiation and depression. *Neural Computation, 13,* 87–111.

Bogartz, R. S. (1990). Evaluating forgetting curves psychologically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 138–148.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition, 20,* 633–642.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33,* 497–505.

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.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9,* 1–7.

Ebbinghaus, H. (1885). *Über das Gedachtnis: Untersuchungen zur Experimentellen Psychologie.* Leipzig, Germany: Duncker & Humblot.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377.

Geinisman, Y. (2000). Structural synaptic modifications associated with hippocampal LTP and behavioral learning. *Cerebral Cortex, 10,* 952–962.

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

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

Glenberg, A. M., & Smith, S. M. (1981). Spacing repetitions and solving problems are not the same. *Journal of Verbal Learning and Verbal Behavior, 20,* 110–119.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review, 7,* 185–207.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 77–99). Oxford, England: Lawrence Erlbaum Associates, Inc.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review, 76,* 82–96.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 883–914.

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.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Nikonenko, I., Jourdain, P., & Muller, D. (2003). Presynaptic remodeling contributes to activity-dependent synaptogenesis. Journal of Neuroscience, *23,* 8498–8505.

Pavlik, P. I., Jr., & Anderson, J. R. (2004, August). *The memory consequences of study after successful recall.* Paper presented at the twenty-sixth annual conference of the Cognitive Science Society, Chicago.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27,* 431–452.

Reed, A. V. (1977). Quantitative prediction of spacing effects in learning. *Journal of Verbal Learning and Verbal Behavior, 16,* 693–698.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288–311.

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning and Verbal Behavior, 17,* 669–680.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103,* 734–760.

Rumelhart, D. E. (1967). *The effects of interpresentation intervals on performance in a continuous paired-associate task* (Tech. Rep. No. 16). Stanford, CA: Stanford University, Institute for Mathematical Studies in Social Sciences.

Scharf, M. T., Woo, N. H., Lattal, K. M., Young, J. Z., Nguyen, P. V., & Abel, T. (2002). Protein synthesis is required for the enhancement of long-term potentiation and long-term memory by spaced training. *Journal of Neurophysiology, 87,* 2770–2777.

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.

Schneider, V. I., Healy, A. F., & Bourne, L. E. J. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46,* 419–440.

Toni, N., Buchs, P. A., Nikonenko, I., Povilaitite, P., Parisi, L., & Muller, D. (2001) Remodeling of synaptic membranes following induction of long-term potentiation. *Journal of Neuroscience, 21,* 6245–6251.

Underwood, B. J. (1969). Some correlates of item repetition in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 8,* 83–94.

VanWinsum-Westra, M. (1990). *Spacing and repetition effects in human memory.* Unpublished doctoral dissertation, University of Nijmegen, The Netherlands.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478.

Wickelgren, W. A. (1973). The long and the short of memory. *Psychological Bulletin, 80,* 425–438.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory and Cognition, 2,* 775–780.

Wickens, T. D. (1999). Measuring the time course of retention. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model* (pp. 245–266). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory and Cognition, 25,* 731–739.

Woodworth, R. S. (1938). *Experimental psychology.* Oxford, England: Holt.

Wu, G., Deisseroth, K., & Tsien, R. W. (2001). Spaced stimuli stabilize MAPK pathway activation and its effects on dendritic morphology. *Nature Neuroscience, 4,* 151–158.

Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology, 8,* 58–81.

## Appendix

Because the math underlying the computation is recursive and a bit complex, the following is an example of the activation and decay computations involved for the session one 14-spacing condition with two tests. This example also includes two tests on the second session. This is a sequence of five presentations, the first being the introductory study and the remaining four being test trials. The long-term retention interval (1 day) is interposed between the third and fourth tests.

The actual average times of this sequence of practice are approximately: 0; 126; 252; 83,855; and 84,888. This indicates that the first study occurs at time 0. The last two times include the 22.5 hr between the end of the first session and the beginning of the second, which is 81,000 sec. To account for the reduced forgetting over this interval we need to multiply the 81,000 sec by the *h* factor (.025), resulting in a scaled value of 2,025. This means that the times after the long interval need to be reduced by 78,975 (the difference between the actual and psychological times) to convert them to a measure of psychological time. This results in the sequence 0; 126; 252; 4,844; 5,877.

Now the activation at each test time can be computed. At Time 126 the first test occurs; according to Equations 4 and 5, $m_1 = ln(126^{-0.177}) = -0.86$. Recall that the decay from the first study ($d_1$) is simply $a$.

To compute the activation of the second test, we need to use the activation at the time of the first test to compute the decay for that presentation. Using Equation 4, $d_2 = ce^{m_1} + a$, which equals 0.27. Since the age of the initial study is now 252, and the age of the first test is 126, from Equation 5 we get $m_2 = ln(252^{-0.177} + 126^{-0.27}) = -0.43$.

To compute the activation of the third test we now need to compute $d_3$, which is $ce^{m_2} + a = 0.32$. The age of the first presentation is now 4,844, the age of the second presentation is 4,717, and the age of the third presentation is 4,591. Therefore activation $m_3$ is $ln(4{,}844^{-0.177} + 4{,}717^{-0.27} + 4{,}591^{-0.32}) = -0.93$.

For the final test in this example sequence we need to compute the decay for the third test (fourth presentation), $d_4 = ce^{m_3} + a = 0.26$. The sequence of ages is 5,877; 5,750; 5,624; and 1,033. Activation $m_4$ is $ln(5{,}877^{-0.177} + 5{,}750^{-0.27} + 5{,}624^{-0.32} + 1{,}033^{-0.26})$, which equals –0.62.