Instrutional Science (2006)
DOI 10.1007/s11251-006-9013-2

Understanding and applying the dynamics of test practice and study practice

PHILIP I. PAVLIK JR.
Pittsburgh Science of Learning Center, Human Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, 15213, USA (e-mail: ppavlik@andrew.cmu.edu)

Received: 27 February 2006; accepted: 9 October 2006

Abstract. Two different methods of practice are available in the learning of simple information, test practice or study practice. Of these two methods of learning, research has generally shown that test practice is superior to study practice. However, this research has not considered the testing advantage with respect to the fact that test learning is uncertain (i.e. if recall fails, nothing appears to be learned) or with respect to the fact that study learning depends greatly on the duration of the study event. The following work clarifies these issues by presenting an integrated computational model of the relative costs and benefits of testing compared to study presentation [based on the ACT-R theory of declarative memory; Anderson, J.R. & Schooler, L.J. (1991). *Psychological Science* 2: 396–408; Pavlik Jr., P.I.. & Anderson, J.R. (2005). *Cognitive Science* 29: 559–586]. This model was applied to determine how test and study practice can be optimally employed to improve learning and provides a framework for understanding the effects of mnemonic strategies in simple memory tasks.

Keywords: distributed practice, efficiency, forgetting, memory, modeling, paired-associate, practice, spacing, study effect, testing effect

Introduction

ACT-R is a large and complex modeling system. Probably the most well known aspect of ACT-R is its production rule system, which describes learning and performance for simple if-then mental procedures. This production rule system operates on both items in the environment (by calling motor or sensory routines if specific conditions are met) and items in memory (by calling for retrieval or encoding of memories). However, the subject of this paper is not this production rule system, but rather the memory items (chunks) that these productions manipulate. The behavior of these memory items is controlled by a memory system (referred to as the declarative memory system in ACT-R), which is characterized by a series of equations that describe the effects of practice and forgetting. These declarative memory equations were first described in Anderson and Schooler (1991), but more recently, Pavlik and Anderson (2005) have extended the model to account for spaced practice of memory items.

However, while this extended ACT-R model captures the effects of spaced practice well, the model does not capture the qualitative or quantitative differences between study practice (seeing a stimulus and possibly thinking about it, but not making an explicit response) and test practice (seeing a stimulus and making an explicit response). The issue of this inaccuracy regarding test practice and study practice differences in the model was first raised by Raiijmakers (2005) in a review of Pavlik and Anderson (2005). In this experiment, participants learned Japanese-English word pairs over the course of several hundred learning trials with 4 levels of repetition (2, 3, 5, or 9 practices), 3 levels of spacing (2, 14, or 98 intervening trials), and 2 retention intervals (1 day or 1 week). Subjects were trained using a procedure in which each pair was introduced with a study-only practice (a presentation of both members of the pair for 5 s). Subsequent spaced practices appeared as test-or-study trials (often referred to as drill practice), which tested items by presenting the Japanese word while waiting for the subject to type the English meaning. These test-or-study trials timed-out after 7 s, and a new 5-s study opportunity was presented after any errors, omissions, or timeouts. This test-or-study procedure was used for the experiment because, according to the assumptions of the Pavlik and Anderson (2005) model, a successful test trial and a study trial result in equivalent learning. This equivalency simplified the modeling since correct retrievals could be treated the same as failures followed by the feedback of a study trial. However, as Pavlik and Anderson (2005) noted, the assumption of equal effect for study trials and test trials was an approximation since a large variety of research has shown advantages for test practice compared to study practice (Allen et al., 1969; Hogan & Kintsch, 1971; Thompson et al., 1978; Runquist, 1983; Slamecka & Katsaiti, 1988; Carrier & Pashler, 1992; Cull, 2000). This paper proposes that the model may be improved by extending it to capture these differences between test practice and study practice.

To extend the model in this way, the experiment completed here compares test practice (both with and without corrective feedback) with study practice. This comparison is suggested by the structure of

ACT-R since these two types of practice represent simple examples of each of the two ways in which an ACT-R model can learn declarative information. The first way is a recall test, in which the context or a cue triggers reconstruction of a piece of information that was encoded at a prior time. The second way is study, in which a piece of information is attended to and a representation is merged into declarative memory. Indeed, it may be true that this distinction is universal for declarative learning and that many if not most declarative learning events in educational tasks could be classified as either test or study type events in that they are primarily generative (result in the reconstruction or reorganization of knowledge from memory) or receptive (result in the construction or organization of knowledge into memory).

While it is clear that both "studying" and "testing" should result in learning, it seems there could be qualitative differences between "studying" and "testing". As a starting point in investigating these differences, this paper assumes that *the amount of learning for each successful recall test is unrelated to the time needed for recall* (this is a standard assumption of the ACT-R architecture and has proven adequate for a variety of models across domains; Anderson & Lebiere, 1998). It is in the context of this default ACT-R assumption about testing that the current paper investigates how best to model study practice.

An accurate model will need to integrate two issues raised by prior research. The first issue is to what extent the amount of learning for a study opportunity is related to the amount of time spent studying. While cognitive psychology has often not varied study durations and simply considered each study trial as a discrete unit of learning (e.g. Carrier & Pashler, 1992; Pavlik & Anderson, 2005), this implicit assumption that all study trials are created equal violates the total time hypothesis (the idea that total-time spent learning determines performance, Cooper & Pantle, 1967). However, despite the probable effect of total time spent learning, the standard spacing effect suggests that the massed practice of a long duration study trial should result in less learning than 2 study trials of equal total time that have had some sort of spacing between them (Dempster, 1989). Because one must reconcile the total-time hypothesis and the spacing effect, a basic hypothesis of the following work is that that there are diminishing marginal returns for learning as study trial duration increases. This hypothesis suggests a reframing of the question "which is better, a test trial or a study trial?", because the answer to this question seems to depend heavily on study duration.

Accepting that study trial effect depends on study trial duration, the new ACT-R model extension proposed here provides a way to describe the diminishing effect of increasing study trial duration on learning for an individual study trial. The experimental data to which this model was fit replicates portions of Metcalfe and Kornell (2003), which also showed this decreasing effectiveness of study trials as study trial duration increases. The following experimental work used study trials with varying durations, either 0 s (a control condition in which no study occurred), 3 s or 7 s. This manipulation provided the data necessary to better determine study effectiveness as a function of time. Zero, 3 and 7 s durations were chosen (rather than even steps such as 0, 3 and 6) because the expectation of decreasing marginal returns implied a larger increment of the manipulation might be necessary to detect differences as study time increased.

A second issue raised by prior work is whether the advantage of test practice relative to study practice is due to an improvement in encoding or a reduction in forgetting (the interaction of practice type and retention interval). While the first option (that practice type affects encoding but not forgetting) may seem more intuitive, manipulations of the spacing effect have been shown to produce less observed forgetting (e.g. Peterson et al., 1963; Fishman et al., 1969; Bahrick, 1979; Pavlik & Anderson, 2005). Indeed, Runquist (1983) proposed that the testing benefit occurs through a decrease in the rate of forgetting. Despite this, other works (e.g. Thompson et al., 1978; Slamecka & Katsaiti, 1988) have given convincing arguments and data suggesting that the benefit for testing does not affect the forgetting rate. To decide this issue the following experiment will use two retention intervals to distinguish forgetting effects from simple encoding effects.

Further, it may be that the forgetting or encoding effect issue hinges on the learning procedures used. Specifically, prior work showing less forgetting for test practice compared a study-only procedure with a test-only procedure (Allen et al., 1969; Hogan & Kintsch, 1971; Runquist, 1983). Unfortunately, it is likely that the test-only procedure (where no feedback occurs) was responsible for the result that forgetting was apparently slower for test practice.

Allen et al. (1969) provides an example of how the test-only procedure can effect conclusions about forgetting rates. In this experiment subjects memorized 27 paired-associates (3 letter nouns – 2 digit number pairings) over the course of a first session of study-only trials (simultaneous paired presentations) and test-only trials (presentations of the noun and recording of the 2 digit response of the subject with no feedback). During this first session subjects were administered either 5 (in the first half or second half of the initial presentation blocks) or 10 initial study-only trials followed by either 0, 1 or 5 test-only trials. There was a final retention test 24 h after this initial session. The basic finding reported was that when test-only trials occurred after the study-only trials, long-term (1-day) memory was enhanced relative to conditions with only study-only trials. For instance, when the five study trials immediately preceded a single test subjects performed significantly better than in the condition with 10 study trials followed by no test trials.

In a study by Runquist (1983) similar results were produced for the memorization of 24 pairs of weakly associated words, all of which received either 1 or 3 study trials, followed 2 min later by a written cued recall test for half the pairs. After the initial practice the full set of words was tested (using a between-subjects design) at one of six retention intervals. The basic result showed an interaction when comparing items only studied with items receiving the cued recall test. This interaction showed that retention decreased more slowly for the tested items as the retention interval increased. Based on this result, Runquist concluded that there must be something special about retrieval that reduces forgetting. He suggests that perhaps recall-testing practice provides contextual learning that applies to further recall test trials or that there was transfer appropriate processing that improved later tests only when practice included testing.

However, even in Allen et al. (1969) an alternative explanation is suggested by their analysis of the conditional portion of errors on trial 1, day 2 given a correct response or an error response on the last trial of day 1. They found that correctness on the last trial of day 1 was highly predictive of correctness on the first trial of day 2 and similarly an error was highly predictive of an error. The key is that unlike study-only practice (or drill practice which provides review in the case of failures), a test-only procedure probably only benefits recallable pairs with each round of practice. This necessarily results in a "rich get richer" and "poor get poorer" scenario since if an item is successfully recalled on a round of testing it is learned more strongly, while if recall fails, nothing is learned and forgetting continues to reduce any further chance of future recall. In contrast, in a study-only procedure, all of the items receive some practice for each trial, so the result is very different. Rather than having two groups of items as in the test-only situation, a study-only procedure will result in items having a well-formed distribution of memory strengths relative to the severely bimodal strength distribution that can result from a test-only procedure. These differences in the strength distributions can cause a very different profile of forgetting after a retention interval. In the case of a test-only procedure, some of the set will demonstrate very slow forgetting (the items gotten correct during learning) while some of the items will not be forgotten because they were never recalled (the items gotten incorrect during learning). This combination will make forgetting seem very slow for the test-only condition. In the case of a study-only procedure, the relatively less variable distribution of strengths will result in faster forgetting.

To avoid this selection problem with test-only procedures, the main comparison in the following experiment will be between a study-only condition and test conditions that include study. As Carrier and Pashler (1992) noted, it is possible to compare a combined test with study procedure against a study-only procedure because one may assume that the learning is equal in each condition for items that are responded to incorrectly in the test condition (since they receive a study trial). This allows the difference between the conditions to be defined by the easier items, which are recalled in the recall-or-restudy condition and studied in study-only condition. When contrasting these conditions, one is therefore looking at the difference between study and testing for items that would have been (or were) recalled.

Further, the inclusion of a test-and-study condition (where study feedback also follows successful recall) along with the test-or-study condition allows the experiment to provide data on the issue of whether or not additional study following a successful recall has any effect. While recent research using a paired-associate procedure (Pashler et al., 2005) has suggested that this study following recall has no effect, the current research will provide a further test of the idea that the extremely short spacing of a study trial immediately after recall results in no perceptible benefit to memory.

Finally, the following experiment looks at the learning and forgetting differences caused by strategy variability. To do this the following experiment will compare subjects trained in a keyword mnemonic procedure with subjects in a free strategy condition. While Atkinson (1975) has shown strong and stable advantages for the sort of mnemonic keyword method analyzed here, Wang and Thomas (1999) have argued that testing must come soon after study presentation for a mnemonic strategy to be durable and effective. Because the experiment here includes a between-subjects manipulation of mnemonics training it may provide data to decide whether there is a significant interaction between mnemonic training and any of the other effects to be investigated such as the amount of learning per trial or the stability of the learning that occurs.

In an effort to form an integrated theoretical argument about these effects, and to explain the results of the following experiment, the analysis of the data gathered relied on an extended formulation of the ACT-R computational modeling system in addition to conventional inferential statistics. One benefit of this modeling analysis was that it allowed the analysis of effects that may be impossible to characterize otherwise. For instance, the modeling analysis allowed a characterization of the difference in the strength of a study trial following a failed test trial compared to a study trial of the same duration that occurred in isolation.

Experiment design

Because Japanese-English vocabulary had been used in the Pavlik and Anderson (2005), it seemed logical to use the same stimulus set to maintain consistency in the effort to create a unified model. The basic task for subjects was to learn Japanese-English vocabulary pairs through study trials and test trials.

The design for the experiment included 32 within-subjects cells each of which began with an initial study-only trial of 5 s. Following the initial study there were two spacing conditions (a spaced trial following either 2 or 30 trials after the initial study trial) crossed with two retention interval conditions (a test-or-study trial either 2 or 60 trials after the spaced trial). For each of these 4 retention interval-by-spacing interval conditions, there were five trial type conditions for the spaced trial (recall-or-restudy, test-and-study, study-only, test-only, or no-practice) crossed with two study duration conditions for the spaced trial types that included study practice (either 3 or 7 s). (Since two of the five spaced trial types did not include study practice, there were eight different study duration-by-practice type conditions. Crossing these eight conditions with the two spacing conditions and two retention intervals yielded 32 cells in the design.)

Each of these cells required either 2 or 3 practice trials to instantiate. For all conditions except the no-practice control condition, 3 trials were needed (one for the initial study, one for the intervening practice trial to define the spacing and practice type manipulations, and one final test-or-study assessment trial to define the retention interval manipulation). In the case of the no-practice control condition, there were only 2 trials required since there was no spaced practice after the spacing interval. This no-practice condition used the same 2 × 2 spacing and retention interval design despite the lack of spaced practice. Because of this, there were 4, 32, 62, or 90 intervening items between the initial study trial and the final test-or-study assessment trial in this condition, corresponding to the sum of the spacing and retention intervals for the spacing-by-retention interval design. The total design for a block therefore required 32 trials for initial study-only trials, 28 trials for trial type and spacing conditions, and 32 trials to assess after the retention intervals, for a total of 92 trials per replication of the 32 cells of the design. This design was replicated once each in two blocks of 160 trials using one randomly selected Japanese-English pair for each replication of each cell. See Figure 1.

This design was instantiated for each subject by randomly ordering the 100 pairs described in the materials section. This pool was then split into 64 pairs which would be used for the experimental conditions and 36 pairs that were used as a pool from which to select buffers. These items were then used to "fill in" a schedule template that was randomly determined for each subject. This schedule template began with an initial block of 20 buffer trials randomized as described below. Following these initial buffers, each of the two experimental blocks was composed of 160 "positions" into which experimental

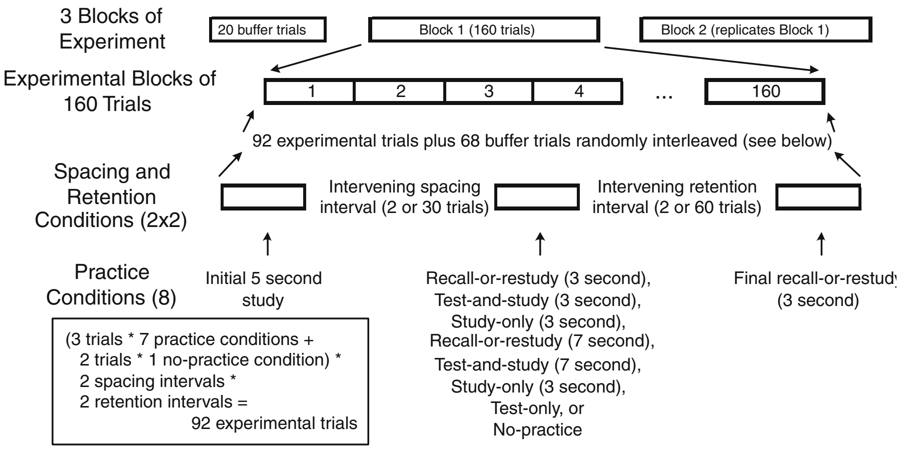
3 Blocks of Experiment
20 buffer trials
Block 1 (160 trials)
Block 2 (replicates Block 1)
Experimental Blocks of 160 Trials
1
2
3
4
...
160
92 experimental trials plus 68 buffer trials randomly interleaved (see below)
Spacing and Retention Conditions (2x2)
Intervening spacing interval (2 or 30 trials)
Intervening retention interval (2 or 60 trials)
Practice Conditions (8)
Initial 5 second study
Recall-or-restudy (3 second), Test-and-study (3 second), Study-only (3 second), Recall-or-restudy (7 second), Test-and-study (7 second), Study-only (3 second), Test-only, or No-practice
Final recall-or-restud
(3 second)
(3 trials * 7 practice conditions + 2 trials * 1 no-practice condition) * 2 spacing intervals * 2 retention intervals = 92 experimental trials

Figure 1. Diagram of experiment design. See text for explanation.

trials could be designated. Distributing the 32 cells composed of 92 trials into these 160 positions was done by first randomizing the 32 cells in a list and then attempting to randomly insert each cell design into the 160 positions. The algorithm to do this performed iteratively for each subject until it found a position for the initial study trial in each cell that resulted in the other 2 (or 1) trials of the cell being both within the block and placed in unfilled positions. The algorithm shuffled the cells together in this fashion for each of the two blocks independently for each subject.

At this point, the schedule template had been filled in with the experimental conditions and the buffer trial positions (the initial 20 trials and 68 unfilled positions in each block) needed to filled for the subject with trials of the 36 buffer items. The most straightforward way to place these trials would be to introduce all of the buffer items in the buffer positions sequentially with study-only trials and then cycle the entire list of buffer items with test-or-study trials. However, this option seemed likely to produce greater overall difficulty early in the experiment and thus possibly skew results. Therefore, a simple algorithm was used to prevent the buffers from biasing the results. This buffer-sequencing algorithm began with a pool of eight randomly selected buffers from the 36 pairs being used for buffers. The buffer trial positions were filled in sequence by random selection from this pool of 8 items. A buffer was retired from this subset if it was used 5 times, and it was replaced with an unused buffer. The first buffer trail for a buffer item was always a 5-s study trial, and the following 4 trials were test-or-study trials with 3-s study intervals. This procedure for interleaving of buffers was intended to insure that the average buffer difficulty was more nearly equal across the two blocks. Since there were 156 buffer trials (20 + 68 + 68), approximately 31 of the buffer items were used 5 times and retired to fill in the buffer positions, but this varied somewhat depending on the random selection from the pool of 8 buffers.

There was also a between-subjects manipulation of mnemonic strategy. For this manipulation, the experimenter read a one-page description of the keyword mnemonic strategy described in detail by Atkinson (1975). Following this short (5-min) reading, the subject was required to briefly summarize (two or three sentences) the keyword mnemonic method to insure that they were paying attention to the experimenter. If the subject failed to summarize the procedure, the experimenter reviewed the sheet until the subject was able to give a brief summary. This procedure advocated finding a keyword similar to the foreign word and then advocated making either a phrase or an image-based link between the keyword and the English response.

Materials

The stimuli and buffers were 100 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 imagability ratings between 343 and 566, with a mean of 464. These ratings were composed according to procedures described in the 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 chosen 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 4 letter English words were used, and 4 to 7 letter Japanese translations were used. Japanese words were presented using English characters (romaji).

Procedures

Participants were scored for motivational purposes, receiving 1 point for each correct response and losing 1 point for each wrong response. Failing to provide a response, either by time-out or providing a blank response, resulted in a 0 score. Participants were paid $9 to $15 depending on their score.

The stimuli were shown on a 19-inch monitor at a resolution of 1024 × 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 sides of the screen. Participant prompts appeared centered horizontally slightly above the vertical center. Participant prompts were in 37 point Tahoma.

All study trials (whether they occurred as feedback or alone) and test trials were cued with the prompts "Study" or "Test" for 0.5 s. Test trials 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 10 s. In the recall-or-restudy condition, a correct response was followed by a 0.5-s presentation of the word "Correct" and the next trial began. An incorrect response in the recall-or-restudy condition was followed by a study trial for the word (which was introduced by the word "Study"). The test-and-study condition was identical to the recall-or-restudy condition if the response was incorrect; however, a correct response was followed by a 0.5-s presentation of the word "Correct" followed by a study trial for the word (which was introduced by the word "Study"). In the test-only condition, no feedback occurred following the test. The study-only and no-practice conditions were self-explanatory.

For the benefit of subjects (to reduce fatigue and improve motivation), the 340 total trials of the experiment were split into blocks of 35 trials (the final block containing 25 trials). To continue on to the next block participants pressed the space bar when they were ready. Few participants paused at these opportunities for rest.

Participants

The experiment used 160 subjects recruited from the Pittsburgh, Pennsylvania community. They were mostly college students responding to an online advertisement. All participants completed the experiment. Eighty participants each were randomly assigned to the two strategy conditions (free strategy or mnemonic training). Sessions lasted slightly less than 1 h. Only participants who professed no knowledge of Japanese were recruited. The participant sample was composed of 83 women and 77 men, with an average age of 21 years. Participant ethnicity was divided as follows: 17 African-American, 57 Asian/Pacific Islander, 73 Caucasian, 4 Hispanic, 7 Middle-Eastern, and 2 Native-American.

Results and discussion

The performance data were aggregated by condition and four repeated measure ANOVAs were completed to examine main effects and interactions in the data. The first ANOVA (retention × spacing × study duration × trial type × strategy condition) compared the recall performance for the test-and-study, recall-or-restudy, and study-only conditions after the retention interval. Main effects were as expected, with retention (*M*s equal 79% and 40% in 2 and 60 trial conditions), spacing (*M*s equal 58.3% and 60.6% in 2 and 30 trial conditions respectively), study duration (57.8% for 3 s and 61.2% for 7 s), and trial type (60.9% for test and study, 63.2% for test or study, and 54.3% for study only) all having significant effects

($F(1, 158) = 1080$, $p < 0.001$, $F(1, 158) = 5.81$, $p < 0.001$, $F(1, 158) = 15.2$, $p < 0.001$, and $F(2, 316) = 30.9$, $p < 0.001$). Effect sizes (Cohen's *d*) also varied by condition with *d*'s equal 2.2, 0.13, 0.20 for retention, spacing and study comparisons respectively. In the case of trial type, the effect sizes for the comparison of the study only condition and the conditions that included testing were 0.36 and 0.47 for the test-and-study and test-or-study conditions respectively. The effect of the strategy condition was also significant, ($F(1, 158) = 21.4$, $p < 0.001$, $d = 0.52$), with *M*s equal to 0.535 in the free strategy condition, and 0.654 in the mnemonics condition.

Of primary importance, this analysis showed no indication of a retention-by-trial-type interaction. As discussed previously, if study practice leads to a less permanent memory encoding, such an interaction with retention interval should occur since forgetting would be faster in the study-only condition when compared to the conditions that included testing (See Figure 2). As shown however, there is no suggestion of quicker forgetting in the study-only condition. Indeed, there is slightly slower forgetting for the study-only trials (though this was not significant). Despite this non-significant result, the 95% CI of the difference between the forgetting effects for testing conditions and the study-only condition was 1.0% to −7.1% indicating that this test was not strong enough to rule out a very slight (1.0%) advantage for test trials. Therefore, the result does establish that any effect of test or study procedures on forgetting is very likely unimportant for instructional application. Further, the study duration by retention interval interaction was not significant, ($F(1, 158) = 2.07$, $p = 0.15$).

The retention interval-by-spacing interaction was also significant with retention after 2 trials with 2 spacing being 1.8% better than

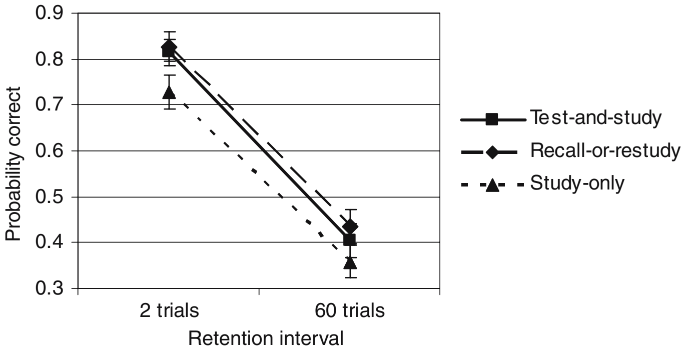

Figure 2. Effect of retention interval depending on trial type. Two *SE* error bars were computed from participant means by condition.

with 30 spacing, while retention after 60 trials with 2 spacing was 6.5% worse than with 30 spacing, ($F(1, 158) = 17.6$, $p < 0.001$, $d = 0.41$). This result is similar to the retention interval-by-spacing crossover-interaction results of Pavlik and Anderson (2005), but may be of a lesser magnitude because the experiment here did not use multiple practices for each item and so could not take advantage of the spacing-by-practice interaction also shown by Pavlik and Anderson.

A second repeated measures ANOVA was identical in design to the first, but only included the study-only condition. Main effects were as expected, with retention (*M*s equal 73% and 36% for 2 and 60 trials), spacing (*M*s equal 52.2% and 56.4% for 2 and 30 trials), and study duration (*M*s equal 51.2% and 57.4% for 3 or 7 s), all having significant effects ($F(1, 158) = 398$, $p < 0.001$, $d = 1.7$, $F(1, 158) = 7.44$, $p < 0.01$, $d = 0.19$, $F(1, 158) = 14.4$, $p < 0.001$, $d = 0.28$). Again, the retention by spacing interval interaction was significant, ($F(1, 158) = 10.8$, $p < 0.01$, $d = 0.44$). There was a 1.9% advantage for 2 trial spacing at a 2 trial retention interval with a 9% advantage for 30 spacing at a 60 retention interval The effect of the strategy condition was also significant, ($F(1, 158) = 17.0$, $p < 0.001$, $d = 0.47$), with *M*s equal to 0.481 in the free strategy condition, and 0.605 in the mnemonics condition.

The ANOVA also showed a 3-way interaction of retention, study duration, and condition, ($F(1, 158) = 7.17$, $p < 0.01$, $d = 0.46$). However, this interaction was not a planned comparison and would not be significant if subject to a correction for multiple comparisons. While not conventionally significant for this reason, this interaction suggests that, when the retention interval is long (60 trials), participants in the mnemonics condition are particularly advantaged during long duration study trials (7 s).

A third repeated measures ANOVA was identical in design to those preceding but included only the test-and-study and recall-or-restudy conditions. In this case only the retention interval (*M*s equal 82% and 42% for the 2 and 60 trial conditions, respectively) and retention-interval-by-spacing (a 1.8% advantage at 2 trials with 2 spacing and a 4.6% advantage at 60 trials for 30 spacing) effects were significant ($F(1, 158) = 872$, $p < 0.001$, $d = 2.2$, $F(1, 158) = 7.48$, $p < 0.01$, $d = 0.22$). The effect of the strategy condition was also significant, ($F(1, 158) = 20.0$, $p < 0.001$, $d = 0.69$), with *M*s equal to 0.562 in the free strategy condition, and 0.679 in the mnemonics condition.

This ANOVA failed to find any advantage for the test-and-study condition, in fact the strong trend ($F(1, 158) = 3.60$, $p = 0.06$) was in favor of the recall-or-restudy condition. This is similar to Pashler et al. (2005), which showed non-significant effects for studying a pair immediately following successful recall of the pair. While in this analysis the spacing effect was not significant ($F(1, 158) = 1.37$, $p = 0.234$), the retention by spacing interaction showed that the spacing effect was still occurring since the significant crossover interaction showed that forgetting was still more rapid after narrowly spaced practice. This result may also imply that the spacing effect is weaker for test practice as compared to study practice. To see if the interaction was significant another ANOVA (retention interval × spacing × study duration × [study trials vs. the average of the two test conditions that included study feedback]) was analyzed; however, the interaction between spacing and trial type was not significant, ($F(1, 158) = 2.31$, $p = 0.13$), similarly the retention by spacing by type interaction was not significant ($F(1, 158) = 1.675$, $p = 0.20$).

A final repeated measures ANOVA (retention interval × spacing × study duration [no study – using test-only condition data, 3 s – using test-and-study and test-or-study conditions, or 7 s – using test-and-study and test-or-study conditions]) was performed to check for an interaction. Figure 3 shows this deceptive interaction, which is very strong ($F(1, 158) = 67.1$, $p < 0.001$) and which could be interpreted as revealing slower forgetting (albeit with less initial learning) in the test-only condition.

Figure 3 seems to suggest slower forgetting for items in the test-only condition, but it actually reveals the effect of item selection on

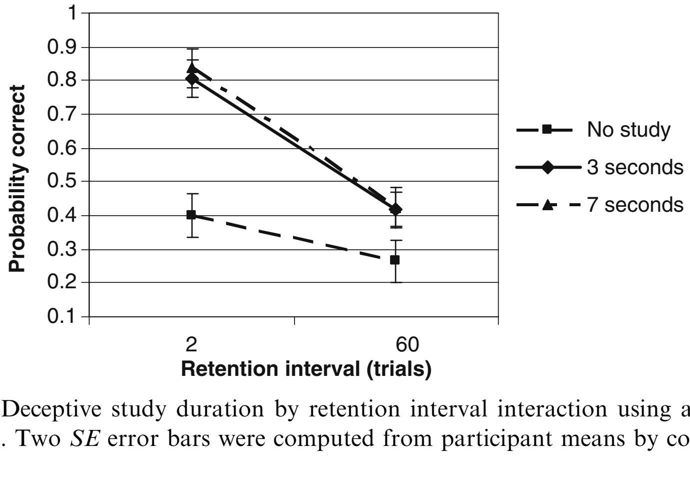

Figure 3. Deceptive study duration by retention interval interaction using a test-only procedure. Two *SE* error bars were computed from participant means by condition.

strength variability in the set of items as discussed in the introduction to this paper. This item selection effect occurred because of the different way retrieval failures were treated. In the test-only (no study) condition, items not recalled get no additional learning and are therefore very likely not to be recalled at final test. In fact, out of the 1280 times subjects received an item in the test only condition (160 subjects × 2 retention intervals × 2 spacing intervals × 2 replications of the design) only 20 (1.56%) were instances where the first recall attempt was a failure and the second recall attempt succeeded. In contrast, 31.7% were instances of two successful recalls, 58.9% were instances of two failures, and 7.73% were instance of a success on the first trial and a failure on the second. This pattern of results supports the notion that there are primarily two groups of items, those recalled successful on the test-only trial and later recalled again after the retention interval, and those for which the test-only trial failed and was followed by failure after the retention interval. Since these groups of items account for greater than 90% of the cases, only a small subset of items (7.7%) that were recalled during testing but not recalled after the retention interval controlled the observed forgetting for the test-only condition in Figure 3.

In contrast, in the test-and-study and test-or-study conditions items not recalled received a study practice after each failure. This meant that a much larger subset of the failed items had the potential to be recalled after a 2 trial retention interval. However, because the of the relative difficulty which had originally caused the failures, recall for these studied items had declined considerably by the 60 trial retention interval. This faster forgetting was not caused by some inherent difference in the stability of information encoded, rather the increased forgetting was observed because the experiment used retrieval failure to effectively select the most difficult items to be part of the subset of items that would be analyzed to determine the forgetting rate for study practice. The addition of study feedback corrected the bimodality of the strength distribution observed for test-only practice, but as a consequence created a pool of items that are not learned well enough to be retained.

Memory model

The advantages of an accurate memory model are twofold. First, the model will be applied to an ongoing project, which uses an algorithm designed from the model to make practice-scheduling decisions

(Pavlik, in press). Since this algorithm must choose between delivering study-only trials or test-or-study trials, having an accurate model of these procedures can be applied to improving the accuracy of scheduling decisions. Second, to the extent that the phenomenon described here may be universal, the general structure of the model may be used to propose theoretical principles that may apply across domains. For instance, the conclusion explores a theoretical basis for how the spacing of practice may interact with the trial type (study-only, test-or-study, test-only).

Since the model used to capture the results is somewhat complex, the following section will build the parts of the model to clarify subsequent discussions of how various hypotheses and explanations agree or disagree with the model. The model used was an extended version of the ACT-R declarative memory theory and captures several major effects in memory. To begin, the standard ACT-R model captures the recency and frequency effects, i.e. that performance is better the more recently or frequently a memory item is practiced (Anderson & Lebiere, 1998). Anderson and Schooler (1991) originally developed this model by showing that memory strength for an item approximates what would be optimal in the environment given the frequency and recency with which an item is encountered. A recent extension of ACT-R (Pavlik & Anderson, 2005) captures the spacing effect. Importantly this extension also captures the spacing by practice interaction (that more practice leads to more effect of spacing) and the spacing by retention interval interaction (that longer retention intervals result in larger spacing effects), effects shown by Underwood (1970) and Fishman et al. (1969).

Activation equation (memory strength function)

These memory effects are captured by an activation equation that represents the strength of an item in memory as the sum of a number of individual memory strengthenings, each of which corresponds to a past practice event (either a memory retrieval or study presentation). Equation 1 proposes that each time an item is practiced, the activation of the item, m_i, receives an increment in strength that decays away as a power function of time (the t_is in Eq. 1 represent the ages of each trial in seconds with n equal to the total number of practices).

To deal with the spacing effect, Pavlik and Anderson (2005) developed an equation in which decay for the ith trial, d_i, is a function of the activation at the time it occurs. This implies that higher activation at the time of a trial will result in the benefit of *that* trial decaying more quickly. On the other hand, if activation is low, decay will proceed more slowly. It is important to note that *every trial has its own d_i* that controls the forgetting of that trial. Specifically, they proposed Eq. 2 to show how the decay rate d_i is calculated for the *i*th trial of an item as a function of the activation $m_{i\text{-}1}$ at the time the trial occurred. Equation 1 shows how the activation m_n after n trials depends on these decay rates, d_is, for the past trials.

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

$$d_i(m_{i-1}) = ce^{m_{i-1}} + a \tag{2}$$

In Eq. 2, c is the decay scale parameter (controlling the amount of decay as a function of activation at the time of learning), and a is the minimum rate of decay. For the first trial of any sequence, $d_1 = a$ since m_0 is equal to negative infinity. These equations are recursive because to calculate any particular m_n one must have previously calculated all prior m_ns to calculate the d_is needed. These equations result in a steady decrease in the long-run retention benefit of each trial when spacing between trials is equal. 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.

Recall equation

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 Eq. 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 given by Eq. 3.

$$p(m) = \frac{1}{1 + e^{\frac{\tau - m}{s}}} \tag{3}$$

In this equation, τ is the threshold parameter and s is the measure of variability in the distribution of performance. The s parameter represents the average noise in activation. When s is high, activation is noisy and recall is less sensitive to changes in activation.

Retrieval time equation

The time to retrieve an item based on its activation in ACT-R is shown in Eq. 4.

$$l(m) = Fe^{-m} + \text{fixed time cost} \tag{4}$$

In Eq. 4, F is the latency scalar parameter, which scales the effect of activation on latency. Fixed time cost usually refers to the fixed time cost of perceptual encoding and motor response processes.

Extension to capture test and study differences

While the ACT-R model above captures the repetition, spacing and recency effects, it cannot distinguish any difference between test trials and study trials. Neither does this model provide an explanation for the benefit of a single study trial as a function of duration. Because the model does not have mechanisms that model these processes, it cannot capture theses effects and will prove incapable of providing an adequate fit for the data collected here. To address this issue, an extension to this model was produced to capture the data. This extension was formulated to capture the fact that the differences in learning for test trials and study trials did not seem to be mediated by changes in the rate of forgetting according to the experiment result.

In this model, each $t_i^{-d_i}$ (see Eq. 1) is multiplied by a parameter (b) that captures the strength of the test or study. Assuming that test trials result in greater learning, this parameter will be greater for successful recalls and less for study trials. Equation 5 shows the how the individual b_is have proportional effect across the span of the retention function.

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

Because of the proportional effect of the b_is, this model expresses the failure to find any significant difference in forgetting between the test and study conditions. Rather, this model is saying that that some learning events are simply stronger than others are. The following section will use this basic characterization of encoding in the model as a framework. This framework allows various ways of computing b_is to be tested.

Five basic models of the effect of study duration were compared. Each of these five study models involved various assumptions about how to compute the b_i weight to scale each $t_i^{-d_i}$ in the activation equation (see Eq. 5). For convenience, each of these study models used the parameters u and v (values in Table 1) to compute the b_i value for a study trial, with the exception of Model 1 which does not use parameters and Model 2 which only requires 1 parameter (u). Further, in all of these models, a correct retrieval trial was fixed at a b_i weighting of 1. These 5 options for determining b_is were fit with 4 different combinations of c (the decay scalar) and/or s (the noise parameter) being optimized in Eqs. 2 and 3. This means that 20 models were developed. All of the models also varied the F (Eq. 4), and s_2, which were needed to find the latency models (the s_2 parameter is described later). Finally, for all of these models, a and τ (Eqs. 2 and 3) were fixed at Pavlik and Anderson (2005) values, and in models where c and s were fixed, these parameters were also fixed at these values (Eqs. 2 and 3). The choice of parameters allows flexibility in capturing the data since using c and s as the main free parameters captures more accurately the slopes of the practice and forgetting functions, while fitting the study function captures initial practice better.

Model 1 tested the standard ACT-R assumption that study trials also have a weight equal to 1. This comparison condition contained no new parameters and is the standard from which the improvement in fit from the additional processes and parameters in the following models can be judged.

Model 2 tested the hypothesis that study weight is constant, but not equal to 1. This model corresponds to the idea that study trials may simply be weaker than test trials and that study duration is not relevant. (In this model the $b = u$ in Eq. 5.)

Model 3 tested the idea that study weight is a linear function of study duration. This idea was tested because it is closely equivalent to one interpretation of how best to account for the benefit of each study trial in ACT-R. Unlike Model 1, in which a single study opportunity occurs whenever study practice is offered to a participant, this model assumes a study trial occurs every 370 ms. The 370 ms figure comes from ACT-R's perceptual motor assumptions and corresponds to an estimate of the minimal time necessary to form an association. While this model is suggested by ACT-R, it is in direct conflict with the spacing effect since it results in no penalty for two back-to-back study trials (e.g. one 740 ms study trial) in comparison to two 370 ms

Table 1. Estimated parameters and model fit statistics for the 5 study models

Study model	Parameters and model statistics								
	c^*	s^*	u	v	F	s_2	χ^2 Recall	χ^2 Latency	Fit statistic
1 only	0.217	0.255	1.000	n/a	1.408	0.686	1157.39	57,717.10	29.13
2 only	0.217	0.255	0.871	n/a	1.297	0.683	510.47	57,561.25	14.41
3 only	0.217	0.255	1.842	n/a	0.800	0.496	606.87	100,000.0	18.68
4 only	0.217	0.255	0.998	0.466	1.267	0.678	389.42	57,026.35	11.63
5 only	0.217	0.255	0.928	1.225	1.307	0.694	450.11	57,671.33	13.04
1 with *s*	0.217	0.419	1.000	n/a	1.414	0.684	457.00	57,715.89	13.21
2 with *s*	0.217	0.364	0.865	n/a	1.289	0.686	165.78	57,557.02	6.58
3 with *s*	0.217	0.372	1.875	n/a	1.277	0.677	193.08	57,241.15	7.19
4 with *s*	0.217	0.350	0.967	0.511	1.280	0.672	132.53	57,055.12	5.80
5 with *s*	0.217	0.362	0.983	0.917	1.233	0.713	129.14	57,876.25	5.76
1 with *c*	0.445	0.255	1.000	n/a	1.207	0.678	429.77	57,434.39	12.57
2 with *c*	0.363	0.255	0.924	n/a	1.208	0.676	305.59	57,260.57	9.74
3 with *c*	0.372	0.255	2.000	n/a	1.188	0.673	333.05	57,068.19	10.36
4 with *c*	0.338	0.255	1.045	0.468	1.198	0.671	245.43	56,843.60	8.35
5 with *c*	0.371	0.255	1.102	0.793	1.221	0.678	240.81	57,482.89	8.28
1 with *c* and *s*	0.401	0.340	1.000	n/a	1.238	0.678	245.77	57,425.22	8.39
2 with *c* and *s*	0.312	0.340	0.901	n/a	1.254	0.667	116.93	57,305.34	5.46
3 with *c* and *s*	0.323	0.342	1.960	n/a	1.214	0.675	134.38	57,056.48	5.84
4 with *c* and *s*	0.297	0.330	0.995	0.521	1.201	0.677	95.723	56,897.56	4.95
5 with *c* and *s*	0.334	0.338	1.205	0.598	1.226	0.686	66.092	57,622.46	4.31

*For each of the 5 study models either, neither, or both *c* and *s* were estimated.

study trials spaced apart. (In this model $b = u$ /10,000* (time in ms/370 ms) in Eq. 5.)

Model 4 captured the study duration effect by proposing that $b = u(1 - e^{(\mathrm{time}-370\,\mathrm{ms})\cdot(-\mathrm{v})/1000})$, where u is the maximum benefit of study and v describes the rate of approach to the maximum. This simple model was intended to capture the fact that the gain from a study trial has diminishing marginal returns and appears to reach an asymptote (Metcalfe & Kornell, 2003). The choice of this function was somewhat arbitrary, since any growth function that displayed a similar pattern of diminishing marginal effect for study as duration increases would have worked as well. For instance, it is likely that a logarithmic growth function would have worked as well. Unfortunately, the experiment was not designed to select amongst such similar study effect models.

Model 5 was formulated and added to the list of hypothetical models because of the misfit of Model 4. Model 5 was designed to capture the fact that a study trial following a failed retrieval had a strong tendency in the model fits to proceed more effectively than a study trial alone (study practice after a success is not so relevant since the high decay in this case wipes out gain). This model tests the assumption that after a failed test feedback-study proceeds more quickly due to prior cue encoding by using two values for v in the equation for Model 4.

Rather than fitting the two values for v, which works only slightly better, Model 5 assumed a simple process explanation to explain the value of v in terms of the stimulus. In this conceptualization, v is divided by the number of terms in the stimulus. This component of the model says that during study trials subjects deploy an attentional resource (typically in a strategic fashion, but also through rote processes) to encode the stimulus being studied. Because this resource is limited, it must be divided among the unencoded components of the stimulus (done in this model by dividing the encoding rate by the stimulus size.) This mechanism captures the idea that the advantage of study after a failed test comes from the opportunity to pre-encode the cue. Because of this, the encoding of the single response term proceeds twice as quickly during the following study opportunity. (While this model fits rather well, the exact function mapping the number of stimulus terms to the value of v remains uncertain and more research will certainly be needed to determine its true form.)

Model fitting and discussion

The following section describes how the parameters were found for the different parameterizations of the model described above. It is important to note that the fit statistic used to optimize the models were not computed in order to do goodness-of-fit testing or likelihood ratio testing. Rather, the fit statistics calculated were simply used as a way to do a quantitative search for different options for representing the data in an ACT-R model.

The fit for the 20 models was computed separately for recall probability and latency. For recall probability (Eq. 3 with each of the 20 model versions above) the model predictions for the conditions were averaged by condition and a χ^2 statistic was computed to describe the goodness of fit to the averages of the condition means by subject as shown in Pavlik and Anderson (2005). On the other hand, the latency models (Eq. 4 with each of the 20 model versions above) were not averaged but compared directly to individual trial latencies for each successful recall in the experiment. Using the ACT-R assumption that latency follows a Weibull distribution as described in Anderson and Lebiere (1998) the latency model goodness of fit was characterized by the summed loglikelihood for the 20459 successful recall latencies which was transformed to a χ^2 distribution by multiplying by -2. As mentioned above, the best fitting F parameter in Eq. 4 and a second s parameter, s_2, were found. While it is possible to fit the overall model with a single s parameter, latency appears to be much noisier than recall probability, while still a reliable function of activation, and thus requires a higher s (this may reflect the fact that there are additional contributions to the latency variability besides variability in memory strengths). In the Weibull distribution, s_2 controls the shape parameter ($1/s_2$) and activation determines the scale parameter. A Weibull distribution is the underlying distribution produced by Eq. 4 assuming noisy activations.

To find the best fitting of these 20 models for both correctness and latency, the models were simultaneously solved for the best fitting F, s_2, study model parameters, and c and/or s to minimize the sum of an overall fit statistic. This fit statistic was computed by dividing the correctness and latency χ^2 values by their degrees of freedom (44 and 20,459) and then taking the sum. While this fitting could have been done successively, i.e. fitting the correctness model by minimizing study model parameter and c and/or s, and then using those parameters to find the best F and s_2, the combined statistic method resulted in a very similar result with less biasing of the fit for correctness results.

In fitting the models, subjects with overall correctness below a cutoff of 1.5 *SD* below the mean correctness were excluded, irrespective of condition. This "cleaning" of the data removed 11 participants. The main reason for removing these participants (who performed near floor) was to prevent them from having too large an effect on efficiency conclusions based on the model, particularly as it pertained to applications of the model. It is important to note that inferential statistics in the results section included these participants since their performance was deemed to reflect real differences due to conditions.

Technical considerations required using the aggregate fit statistic. The primary problem was that the study model needed to be fit at the level of the conditions because otherwise it would be biased by the fact that more feedback study trials occur for difficult items. Given this, if one estimated parameters for study effect, these parameters would also capture some part of individual and item differences. An aggregate fit by condition may have been avoided by a simultaneous fit of both study effect and individual and item differences, but this would have involved a simultaneous fit of 1000s of parameters and was therefore impractical. While correctness therefore needed to be fit by condition, correct latency had variable frequency by condition and did not seem as amenable to a similar condition based fit. The fit statistic was a practical compromise to deal with these issues.

Given the final fit of the models (Table 1), several things are clear. First, it seems that c and s are the minimum number of recall probability model parameters that need to be estimated (despite the fact that s alone is considerably better than c alone). One reason why the Pavlik and Anderson (2005) default values do not work here is likely because of the differences in trial delivery. Specifically, in Pavlik and Anderson, inter-trial intervals were between 1 and 2 s, while for the current experiment they were only 0.5 s. It is likely that by having these shorter intervals learning became more variable for each trial, requiring a higher s. Similarly, the increase in c might be due to an overall greater effect of spaced practice, which might occur due to generally greater interference with shorter inter-trial times. Interference has been shown to increase spacing effects (Bjork & Allen, 1970).

Figure 4 shows a graph of the final five study (b) models from Table 1. These models describe the effect of study duration on encoding strength. Of these models, Model 5 fit the data best.

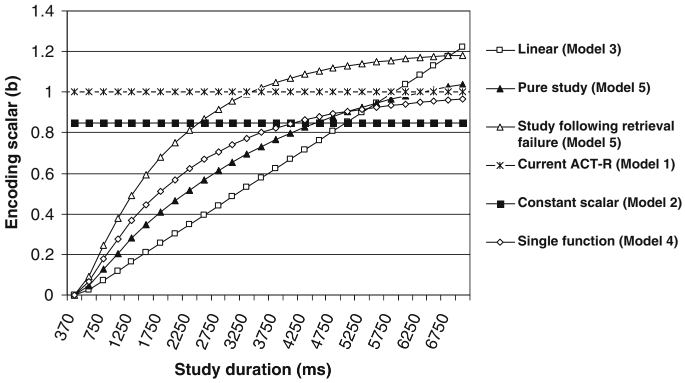

Figure 4. Model predictions for effect of study time on the encoding strength scalar parameter.

The overall best fitting recall probability model and human data are shown in Figure 5, which includes all 44 data points modeled. The fit is clearly quite good. The bottom line of Table 1 gives the values for the 6 free parameters estimated to fit the data. While the model fitting procedures do not allow statistical testing, Table 1 shows the relatively large improvements of fit using Model 4 and especially Model 5.

The recall latency model had a proportionally worse fit, likely due to the large amount of noise in the data. Because of this noise, the s_2 parameter, used to fit the distribution of latencies (given the Weibull assumptions described above), was much higher than the noise parameter used in the recall equation (Eq. 3). This model reflected a raw correlation of -0.33 between activation and correct response latency. Mean correct response latency (first keystroke) was 2.267 s. As ACT-R assumes, there appeared to be no correlation between activation of a memory and retrieval failure. In fact, the correlation was merely 0.02 between failure latency and activation. Mean failure latency was 5.195 s.

The following discussion focuses in more depth on the issues surrounding stimulus encoding and on how it is explained by the model. The first aim of this discussion is to reveal how the study duration model can be used to explain the effect of the keyword strategy taught to subjects. A second section describes the practical implication of the current paper for determining the optimally efficient study trial duration given the model.

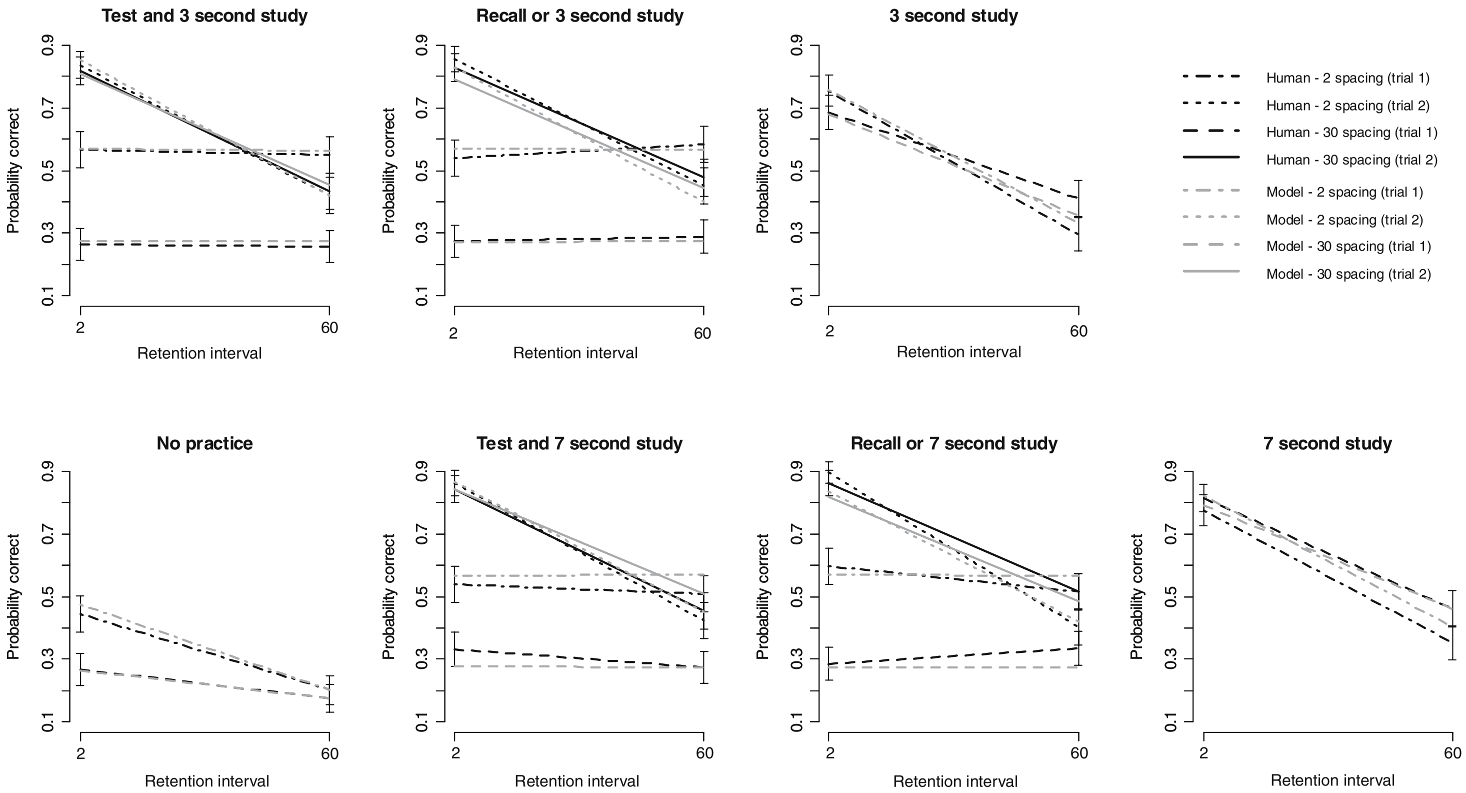

Figure 5. Human and model correctness data for the experiment. Two *SE* error bars were computed from participant means by condition.

Mnemonics (strategy use)

One-half of the participants had the keyword mnemonic technique described to them and were encouraged to use it. This condition was included in the experiment because the keyword mnemonic technique has a history of providing benefits in tasks such as this one. One of the most comprehensive examples of the positive effects of experimenter-supplied keyword mnemonics is Atkinson (1975). In this review of his work, he discusses an experiment (Atkinson & Raugh, 1975) in which participants learned 120 Russian-English word pairs over three days with experimenter provided keywords or without. This experiment showed a recall level of 72% in the mnemonics condition the day after the final learning session with only 46% recall in the control condition. This benefit was also found to be very stable, with keyword condition recall at 43% and control condition recall at only 28% after 6 weeks. The results of the experiment here generally confirm the findings of Atkinson and Raugh and extend those findings to a paradigm where mnemonics are not supplied by the experimenter, but rather must be created by the participant in the experiment.

Two models were made to capture the two strategy conditions. For these two models, the parameters for c (decay scalar) and s (correctness noise) determined above were accepted and the best fitting F (latency scalar), s_2 (latency noise),v (acquisition rate parameter) and u (maximum acquisition) for each strategy were found by minimizing fit statistics composed as describe previously for the full model. These fits included all 80 subjects in each condition.

Since the new parameters described study functions, these functions are compared in Figure 6. Figure 6 reflects v values = 0.434 and 0.988, with u values = 1.52 and 0.867, for mnemonic and free strategy conditions respectively.

As expected, mnemonic strategy parameters differed from the free strategy parameters considerably. Since a keyword mnemonic typically involves a somewhat complex process including devising a keyword, setting a scene and/or saying a sentence, it makes sense that the benefit of mnemonic practice rises less steeply than for the free strategy condition. It also seemed that mnemonics provided an overall improvement in the asymptotic strength of an encoding.

Figure 7 shows the fit of these strategy models to the human data for the study-only condition (including the condition where no-practice occurs after the spacing interval as a control), aggregated across

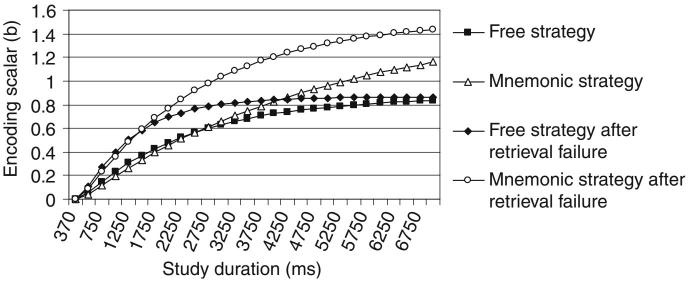

Figure 6. Encoding scalar model by strategy condition for failure study trials and study-only trials.

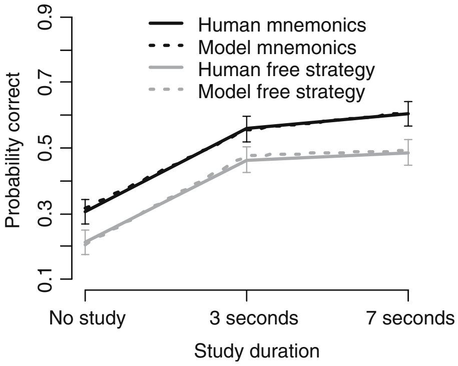

Figure 7. Graph of study-only and no study conditions by strategy condition for human subjects and model. Two *SE* error bars were computed from participant means by condition.

the spacing and retention interval conditions. The points in Figure 7 show the average of an overall model that was fit by strategy condition to the same 44 conditions as Figure 5. As can be seen in Figure 7, the fit is excellent, and it mirrors what might be expected from the functions in Figure 6. In the case of mnemonics, the recall from the initial study results in superior recall for the no study condition. This advantage for mnemonics is preserved in the case of both the 3 and 7-s study conditions.

It is interesting to note that the *F* parameter (the scalar that controls recall latency magnitude) also varied in a meaningful way by strategy. *F* was greater for the mnemonics strategy condition (*F* equaled 1.30 in the mnemonic condition and 1.12 in the free condition), suggesting the need for strategy use at retrieval may have slowed performance. While the latency comparison for the data between conditions was not significant, ($F(1, 159) = 2.08$, $p = 0.151$, two-tailed test), the model and data both suggest a trend toward slower recall when using mnemonic strategies. This would agree with work by Crutcher and Ericsson (2000) who have shown that when using a dual-task interference paradigm there appears to be strategy use at retrieval when encoding has involved the keyword method.

Efficiency implications

Given the functions in Figures 4 and 6, it becomes possible to optimize the study trial unit task so that it is maximally efficient in terms of the study model. Optimization requires consideration of two cost factors, fixed time costs and variable time costs. Since the unit task always contains a 500 ms pre-trial prompt, this must be included in the fixed costs. This makes the *total time cost* = 500 ms + *time of trial*. Using this cost value and the value of *b* from Model 5, Figure 8 can be graphed to show the efficiency of various study durations given the 3 parameter sets (overall, mnemonic and free strategy) for Model 5.

The curves in Figure 8 were surprising because they implied a much shorter optimal study time than was used in Pavlik (in press). Each graph shows the average efficiency of a trial (trial gain in terms of the study model *b* value/trial cost in milliseconds) given the total

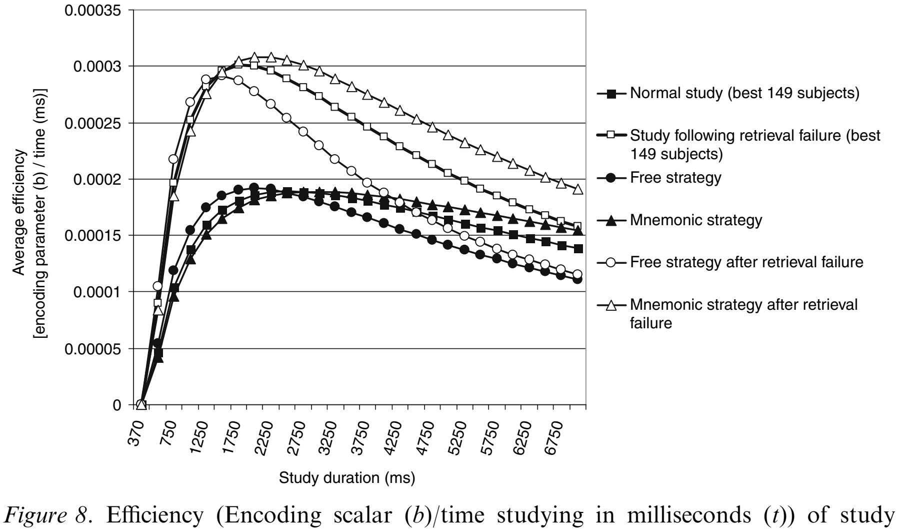

Figure 8. Efficiency (Encoding scalar (*b*)/time studying in milliseconds (*t*)) of study trials depending on strategy condition and test result, and for the overall model in Figure 5.

duration of the unit task (including the 500 ms prompt). The x-axis does not include the 500 ms pre-trial prompt, only indexing the duration of the trial, so the functions begin rising at 370 ms (the fixed minimum time for the perceptual components to encode the stimulus in ACT-R). The graphs rise at first because initial learning is quick. When learning begins to slow as encoding gets closer to u, a point is reached where so little is gained that the average efficiency no longer increases with additional study time, but rather decreases.

Therefore, the maximum on each graph predicts optimal study duration efficiency. The existence of this maximum has obvious implications for improving learning performance. As might be expected, the graphs show that people instructed in mnemonics (presumably more time consuming than a less constrained learning procedure) have later optimums then the free strategy group. This shows one interesting implication of these graphs, which is that the optimum study time depends on the strategy invoked. Table 2 shows these optimal durations.

These different optimums by strategy suggest an interesting confound in many researchers work, which is that the study duration interacts with the efficiency of the learning task. Typically, a comparison between two learning procedures uses trials of equal duration for each strategy. Because of the interaction between strategy efficiency and study duration, however, a better comparison of mnemonics and a free strategy condition would use the optimal trial durations for each strategy with a fixed total period of learning with each strategy. This type of comparison would avoid the implicit penalty incurred when the study duration is not optimized for the strategy condition. Based on this analysis, using constant durations could be biased depending on the optimal study durations. For instance, in Figure 8 above, one could select study durations greater than 1.8 s and have

Table 2. Model prediction of optimal study duration by condition

Modeled condition	Optimal study duration (seconds)
Normal study (best 149 subjects)	3.1
Study following retrieval failure (best 149 subjects)	2.5
Free strategy	2.6
Mnemonic strategy	3.5
Free strategy after retrieval failure	2.2
Mnemonic strategy after retrieval failure	2.7

mnemonics be favored, or 1.8 s or less and favor a free strategy. The root of this prediction can be seen in Figure 6 where over the early portion of the function one can see that the faster encoding of a free strategy briefly overwhelms the greater potential of the mnemonic strategy. The implication is that the success of mnemonics in the current experiment was partially a function of the study duration and if study times had been very short one would have seen less of a benefit for mnemonics or even an advantage for a free strategy.

General discussion

The integrated model presented here has practical and theoretical implications. For example, in Pavlik (in press) subjects practiced the same set of Japanese-English pairs practiced in this experiment. The experiment used a single 1-h first session of practice trials followed by a retention test the following day. This retention test assessed the effect of a wide-spacing condition (where the set of 100 pairs was introduced one by one with study-only trials, and then the same sequence was cycled with test-or-study trials until the hour was concluded) and an "optimized" condition (where an item selection algorithm choose items for practice to maximize expected gain/ expected cost for each trial according to the model, resulting in trials for items being interleaved such that individual items had gradually increasing spacing across a series of repetitions). While Pavlik showed a significant advantage for this "optimized" efficiency-based selection without the new model mechanisms proposed in this paper, adding the new mechanisms to the model should allow further gains in practice efficiency at two points in the Pavlik practice-scheduling algorithm.

The first decision that can be improved is the duration of each study trial. The model described in this paper allows this improvement by providing information about the amount of study trial benefit as a function of time. This information about the study function allows a prediction of the most efficient duration of practice as was shown. The practical significance of using this efficient duration is not to be overlooked since over the course of 100s of learning trials, a 1 or 2 s reduction in time per study trial (with only a slight reduction in learning) adds up to a considerable advantage for using the shorter duration.

The second decision that can be improved is the choice of whether to give a test-or-study trial of an item or a study trial for any particular trial. This decision is improved because the model now allows us to compute and compare the relative benefits from study trials and test trials. Since the practice-scheduling algorithm in Pavlik (in press) balances the chance of failure during drill practice (which results in study practice causing learning) and the chance of success (which results in test practice causing learning) to determine the expected learning for a trial, greater accuracy in capturing the difference between test trials and study trials will result in more accurate practice scheduling.

Besides these specific improvements in local practice scheduling decisions, this work also has general implications for instruction. First, because failed test trials alone appeared to cause no learning, one can conclude that learning from test-only trials depends on ability to recall. This point means that generally speaking, the efficiency difference between test practice and study practice depends on current learning of the item. Test practice can only be superior when current learning allows retrieval.

The implication is similar for test trials that give study trial feedback if failure occurs. In this case, failure does not have the same catastrophic effect on learning since the feedback study provides fallback learning. However, despite this fallback study, it was still found that frequent retrieval failures result in poor learning efficiency (Pavlik, in press). In this case, the poor efficiency was not due to failure to learn, but rather it was due to the large time cost of failure (e.g. in the experiment here failure latency was more than twice success latency).

Because of these large costs for failure to recall, it appears that a procedure of gradually increasing spacing for each repetition of an item (Pimsleur, 1967; Landauer & Bjork, 1978; Cull et al., 1996; Pavlik, in press) should be best for both test-only procedures and for testing procedures that allow additional study if testing fails. In both cases, gradually increasing spacing results in better learning since retrieval can be maintained with a high probability, preventing the consequences of retrieval failure. In contrast, study-only procedures should have a much wider optimum spacing since they do not require retrieval and so are not associated with either consequence of failed retrieval (failure to learn or long latency).

Not only do the results support these practical implications for instruction, but they also suggest a theoretical perspective on why differences in learning procedures cause differences in forgetting and encoding. The data here provide some support for the idea that forgetting does not change as a function of the encoding procedure (whether test, study or mnemonic). In fact, manipulations of encoding typically seem to affect recall but have never been convincingly shown to effect forgetting rate per se (e.g. Bentin et al., 1998, show approximately the same rate of forgetting for different encoding tasks; McBride & Dosher, 1997, found no forgetting differences for explicit and implicit memory). While results by Wang and Thomas (1995) show faster forgetting for a mnemonic condition as compared to a contextual association condition (reading words in sentences with definitions), one might question these differences in forgetting for the same strength variability reason as in the test-only paradigm. It seems plausible that forgetting differences in the Wang and Thomas procedure were created by differences in strength variability caused by difference in the encoding strategies used in their contextual and mnemonic encoding conditions. Specifically, if the contextual encoding condition resulted in more strength variability in learning per trial than the mnemonic condition, the contextual condition may have showed seemingly slower forgetting. This explanation is made more likely by the fact that mnemonic keywords were provided in this study, so variance in that condition would not have been influenced by the variable difficulty of finding keywords. Wang and Thomas also observed that repetition reduced the forgetting differences between the mnemonics and context conditions. Since repetition should result in less variability in strength (reduction in the standard error of the mean strength for each pair as the number of trials accumulates for each pair) the argument that forgetting difference in this case were caused by variability in strength is further supported.

Therefore, it seems plausible that encoding related forgetting effects in the literature may be explainable as strength variability effects (e.g. Runquist, 1983; Wang & Thomas, 1995). This agrees with the theory proposed here because in the model encoding processes and duration of encoding determine the strength of the encoding, not the forgetting rate. In the model, it is supposed that this encoding occurs in working memory, which means the strength of each b in Eq. 5 is determined by working memory ability and processes in working memory.[1] Because of this, the model corresponds well with work such as Kane and Engle (2000) and Baddeley et al. (1984), in which encoding processes occur in working memory and are susceptible to dual-task interference.[2] The key point here is that this work leads to a direct reason why encoding manipulations (e.g. differences in encoding strategy) may effect encoding strength (because they change the strength of the original instantiation in working memory before forgetting becomes a factor), but very little reason to explain why they would effect forgetting rates except by causing variability in strength among items.

So what makes the spacing effect different so that the model here characterizes it by differences in forgetting rates? In the example encoding manipulations for which reliable forgetting effects cannot be found (a literature search turns up a surprising paucity of investigations), the parameters of the manipulations do not depend on prior practice directly. The spacing effect on the other hand depends directly on prior practice dynamics. One might speculate that this prior learning is the cause of the change in the forgetting rate for new repetitions that appears to occur as a function of spacing. A neurally based argument for why the characteristics of long-term potentiation (LTP) correspond to the characteristics of the spacing model was offered in Pavlik and Anderson (2005). According to this argument, new repetitions are stable in the face of interference only to the degree that LTP processes have had time to recuperate from prior learning events. In contrast, when prior learning is controlled, forgetting may be independent of the way material is learned. While this explanation does not directly address contextual or encoding variability and how they may affect forgetting (Glenberg, 1976; Martin, 1968) in a way similar to spacing, it is clear that contextual or encoding variability are always relative to prior practice, so again, it seems that the forgetting rate for new learning is controlled by an interaction with prior learning.

Notes

1. Indeed, while the experiment here did not deal with individual or item differences, an expanded model dealing with these differences (Pavlik, in press) captures these differences in an equivalent fashion, implying that these individual differences occur in working memory and are essentially encoding effects.
2. The model could capture this sort of dual-task encoding interference in a variety of ways, such as penalizing the rise parameter (v) or by reducing the effective duration of the study to account for the dual-task performance.

Acknowledgments

This research was supported by NIH training grant MH 62011, NIMH R01 MH 68234 and by Ronald Zdrojkowski.

References

Allen, G.A., Mahler, W.A. & Estes, W.K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior* 8: 463–470.

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

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

Atkinson, R.C. (1975). Mnemotechnics in 2nd-language learning. *American Psychologist* 30: 821–828.

Atkinson, R.C. & Raugh, M.R. (1975). An application of the mnemonic keyword method to the acquisition of Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory* 104: 126–133.

Baddeley, A., Lewis, V., Eldridge, M. & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General* 113: 518–540.

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

Bentin, S., Moscovitch, M. & Nirhod, O. (1998). Levels of processing and selective attention effects on encoding in memory. *Acta Psychologica* 98: 311–341.

Bjork, R.A. & Allen, T.W. (1970). The spacing effect: Consolidation or differential encoding. *Journal of Verbal Learning & Verbal Behavior* 9: 567–572.

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

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

Cooper, E.H. & Pantle, A.J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin* 68(4): 221–234.

Crutcher, R.J. & Ericsson, K.A. (2000). The role of mediators in memory retrieval as a function of practice: Controlled mediation to direct access. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 1297–1317.

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

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

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

Fishman, E.J., Keller, L. & Atkinson, R.C. (1969). Massed versus distributed practice in computerized spelling drills. In R. Atkinson & H.A. Wilson, eds, *Computer assisted instruction*. New York: Academic Press.

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.

Hogan, R.M. & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior* 10: 562–567.

Kane, M.J. & Engle, R.W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 336–358.

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

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: An encoding variability hypothesis. *Psychological Review* 75: 421–441.

McBride, D.M. & Dosher, B.A. (1997). A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General* 126: 371–392.

Metcalfe, J. & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General* 132: 530–542.

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

Pavlik Jr., P.I. (in press). Timing is an order: Modeling order effects in the learning of information. In F.E. Ritter, J. Nerb, T. O'Shea & E. Lehtinen, eds, *In order to learn: how ordering effects in machine learning illuminate human learning and vice versa*. New York, NY: Oxford University Press.

Pavlik Jr., P.I. & Anderson, J.R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science* 29: 559–586.

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

Pimsleur, P. (1967). A memory schedule. *Modern Language Journal* 51: 73–75.

Raiijmakers, J.W. (2005). [Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect.]. *Cognitive Science* 29: 559–586.

Runquist, W. (1983). Some effects of remembering on forgetting. *Memory and Cognition* 11: 641–650.

Slamecka, N.J. & Katsaiti, L.T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 716–727.

Thompson, C.P., Wenger, S.K. & Bartling, C.A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory* 4: 210–221.

Underwood, B.J. (1970). A breakdown of the total-time law in free-recall learning. *Journal of Verbal Learning & Verbal Behavior* 9: 573–580.

Wang, A.Y. & Thomas, M.H. (1995). Effect of keywords on long-long retention: Help or hindrance?. *Journal of Educational Psychology* 87: 468–475.

Wang, A.Y. & Thomas, M.H. (1999). In defense of keyword experiments: A reply to Gruneberg's commentary. *Applied Cognitive Psychology* 13: 283–287.