Mechanisms underlying the Spacing Effect in learning: A comparison of three computational models

Matthew M. Walsh[1], Kevin A. Gluck[2], Glenn Gunzelmann[2], Tiffany Jastrzembski[2], Michael Krusmark[2,3], Jay I. Myung[4], Mark A. Pitt[4], and Ran Zhou[4]

[1] The RAND Corporation, Santa Monica, CA

[2] Air Force Research Laboratory, Wright-Patterson Air Force Base, OH

[3] L-3 Communications, Wright-Patterson Air Force Base, OH

[4] The Ohio State University, Columbus, OH

*Authorship is alphabetical except for the first author.

January 2018

(*Journal of Experimental Psychology: General*, in-press)

Send correspondence to Matthew Walsh

Phone: (973) 978-4992

Email: mmw188@gmail.com

Abstract

The spacing effect is one of the most widely replicated results in experimental psychology: Separating practice repetitions by a delay slows learning but enhances retention. The current study tested the suitability of the underlying, explanatory mechanism in three computational models of the spacing effect. The relearning of forgotten material was measured, as the models differ in their predictions of how the initial study conditions should affect relearning. Participants learned Japanese-English paired associates presented in a massed or spaced manner during an acquisition phase. They were tested on the pairs after retention intervals ranging from 1 to 21 days. Corrective feedback was given during retention tests to enable relearning. The results of two experiments showed that spacing slowed learning during the acquisition phase, increased retention at the start of tests, and accelerated relearning during tests. Of the three models, only one, the Predictive Performance Equation (PPE), was consistent with the finding of spacing-accelerated relearning. The implications of these results for learning theory and educational practice are discussed.

Keywords: spacing effect, relearning, Predictive Performance Equation, ACT-R, education

1. Introduction

The acquisition and retention of knowledge is impacted by many factors. Of these, one of the most intensively studied is the temporal distribution of practice. Numerous experiments have shown that separating practice repetitions by a delay slows learning but enhances retention (for recent reviews, see Benjamin & Tullis, 2010; Cepeda et al., 2006; Delaney, Verkoeijen, & Spirgel, 2010). This is called the *spacing effect*. The spacing effect is of great theoretical interest because how spacing impacts acquisition and retention can constrain theories and computational models of learning and memory (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009; Pavlik & Anderson, 2005; Raaijmakers, 2003). The spacing effect is also educationally relevant because understanding how spacing affects memory should lead to the design of study and training schedules that increase the efficiency of learning and the maintenance of mastery (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, *submitted*).

In this paper, we evaluate the adequacy of the core computational mechanisms in three computational models of the spacing effect. The first extends the activation equations from Adaptive Control of Thought-Rational (Pavlik & Anderson, 2005), the second builds on Anderson & Schunn's (2000) General Performance Equation (Jastrzembski & Gluck, 2009; Walsh et al., *submitted*), and the third generalizes the Search of Associative Memory model (Raaijmakers, 2003). Although there are other computational models of the spacing effect (e.g., Benjamin & Tullis, 2010; Estes, 1955; Mozer et al., 2009), we focused on these three because they are among the most widely studied and used, and they contain distinct mechanisms that explain the spacing effect.

The three models account for key findings related to how the temporal distribution of practice affects knowledge acquisition and retention. For example, all can explain three key findings in the literature (Pavlik & Anderson, 2005; Raaijmakers, 2003; Walsh et al., *submitted*). First, spaced study slows initial acquisition. Second, spaced study enhances retention. Third, including more space between study repetitions improves retention to a point, after which it has diminishing or negative effects.

Notwithstanding these similarities, the models make different predictions regarding whether spacing affects subsequent relearning. However, few studies have examined relearning, and almost none has examined relearning following spaced practice. This is somewhat surprising given that in Ebbinghaus's pioneering work on human memory, he used a measure of savings during relearning to quantify the effect of spaced rehearsal on memory strength (1885/1964; for one contemporary exception, see Pavlik & Anderson, 2005). To address this gap, we conducted an experiment to investigate how spacing impacts relearning. We asked whether spaced practice, delivered during an initial study session, affects relearning following long retention intervals. As discussed throughout this paper, the results provide new theoretical insight into the dynamics of learning and memory, and they have implications for the scheduling of learning events in education and training.

1.1 The Spacing Effect

The temporal distribution of practice affects learning and retention (Ebbinghaus, 1885/1964; Jost, 1897; Thorndike, 1912). This is seen in experiments that vary the elapsed time between item repetitions. When an item is studied repeatedly without interruption, learning is *massed*. Alternatively, when repetitions are separated by intervening items or time, learning is *spaced*. In a variation of this procedure, all practice is spaced, but the inter-trial interval (ITI) separating repetitions is longer for some items than for others. The canonical finding is that greater spacing during study slows learning but enhances retention. For example, in a study by Greeno (1964), participants memorized digit-word pairs (Table 1). Each pair appeared three times during a training series, after which a test was administered. Some pairs repeated after 0 or 1 intervening items during the training series, and some repeated after an average of 15 intervening items. The proportion of correct responses on training trials 2 and 3 was higher for massed versus spaced repetitions, whereas the proportion of pairs correctly recalled during the retention test was higher for spaced versus massed repetitions (Table 1).

Table 1. Results from Experiment 1 of Greeno, 1964

	Proportion of Items Correct	
Condition	Training Trials 2 and 3	Test Trial
Spaced (15 items separating repetitions)	46%	66%
Massed (0 or 1 item separating repetitions)	92%	53%

The spacing effect is extremely general. It occurs in tasks that involve declarative knowledge (Hintzman & Rogers, 1973; Janiszewski, Noel, & Sawyer, 2003), procedural skills (Lee & Genovese, 1988; Moulton et al., 2006), and academic competencies (Rohrer & Taylor, 2006; Seabrook, Brown, & Solity, 2005). The spacing effect has been replicated in laboratory studies (Cepeda et al., 2006), and in ecologically-valid educational settings (Carpenter et al., 2012). Children and adults show spacing effects, as do different animal species (Delaney, Verkoeijen, & Spirgel, 2010). Its ubiquity suggests that it reflects the operation of general memory mechanisms not tied to modality or stimulus.

Researchers have also manipulated the amount of time between study opportunities – that is, the ITIs – to determine the level(s) of spacing that optimize retention. In these experiments, the amount of time between study opportunities varies by condition, and performance is measured after a fixed retention interval (RI) following final study. Some amount of spacing usually improves performance. However, the duration of the ITI has a non-monotonic effect on retention. As the ITI increases, final retention rises sharply before peaking, after which it gradually declines. For example, in a study by Young (1971), participants memorized digit-trigram pairs. Each pair appeared twice, with the repetitions separated by ITIs ranging from 0 to 17 trials. All items were tested 10 trials after the second repetition. The proportion of correct responses on the final test trial was lowest for items presented at an ITI of 0 trials (41% correct), highest for items presented at an ITI of 7 trials (53% correct), and somewhat lower for items presented at an ITI of 17 trials (45% correct). The effects of ITI are modulated by the duration of the retention interval (RI). As the RI increases, so too does the length of the ITI that maximizes retention (Cepeda, Vul, Rohrer, Wixted & Pashler, 2008; Glenberg, 1976). This has been referred to as the "temporal ridgeline of optimal retention" (Cepeda et al., 2008).

1.2 The effect of spacing on relearning – An empirical gap

Although hundreds of spacing effects studies have been published, virtually none has examined the effects of spacing on relearning (for one exception, see Pavlik & Anderson, 2005). The near absence of such studies reflects two general trends in research on the spacing effect. First, most experiments use a single-session design where items are studied repeatedly and tested within minutes (Greeno, 1964; for a review, see Cepeda et al., 2006). Relatively little forgetting occurs across such short timescales, limiting the potential to observe differences in relearning. Recent experiments that have used longer retention intervals provide support for the utility of spaced study, but they also reveal the boundaries of its effectiveness. Although spacing reliably improves retention, memory performance after multiple days is low regardless of whether practice is massed or spaced (Pashler, Zarow, & Triplett 2003; Sobel, Cepeda, & Kapler, 2011; Vaughn, Dunlosky, & Rawson, 2016). At first blush, these results suggest there may be little reason to bother with spaced learning when the anticipated retention intervals are greater than 1 day. Across such intervals, the benefits of spacing in absolute terms are quite small. Although spacing improves retention, decay inevitably drives memory back toward zero.

Second, in most spacing effect experiments, memory is evaluated using a single recall test. As a performance measure, recall is limited in that it only provides information about whether the accessibility of an item in memory is above or below the threshold for successful retrieval (Nelson, 1985). The fact that an item can no longer be retrieved does not mean that it has been lost from memory. With additional exposures, the item may again become accessible. Indeed, relearning of forgotten items is considerably faster than original learning (Ebbinghaus, 1885/1864; Groninger & Groninger, 1980; Macleod, 1988; Nelson, 1978). Measures based on savings during relearning may be more sensitive for detecting small amounts of information that remain in memory after multi-day retention intervals in spacing effect experiments.

This empirical gap leaves open the question of whether and how spacing, delivered during initial study, affects subsequent relearning. If initial spacing increases an item's memory strength, as indicated by greater retention following spaced versus massed practice, it may also accelerate subsequent relearning. Another possibility is that following long enough retention intervals, after which memory performance drops to floor, the effects of initial spacing may no longer affect performance. A final possibility is that initial spacing may impact retention and relearning in opposing ways, by enhancing retention and hindering relearning. As a conceptual realization of this hypothesis, spacing does impact initial learning and retention in opposing ways.

The answer to this question is relevant to descriptive and computational accounts of the spacing effect, as we explore in the following sections. The answer is also relevant to educational applications of spacing effect research. Spacing enhances retention, but after long enough retention intervals, memory performance is low regardless of how items were initially studied. With this in mind, ease of relearning becomes an equally important criterion for judging the educational value of an intervention. Durable learning is important, but so too is efficient relearning (Rawson & Dunlosky, 2011). Nelson (1985) noted this, saying

> *Educators generally realize that most of what they teach their students does not remain recallable, or even recognizable, for extended periods of time. Rather the hope is that their teaching will have been beneficial because the student could relearn the information relatively quickly… How should information be taught so as to maximize these benefits? (p. 478)*

If the manner in which items were initially studied accelerated subsequent relearning, even after the items were completely forgotten, that would demonstrate the educational value of spacing.

1.3. Descriptive theories of the spacing effect

Several theories have been proposed to explain the spacing effect. They can be divided into two categories, descriptive and computational, depending on their level of specificity. Descriptive theories provide general accounts of how spacing impacts knowledge acquisition and retention, but stop short of implementing accounts computationally. Descriptive theories include the deficient processing hypothesis, the context variability hypothesis, and the study-phase retrieval hypothesis. According to the *deficient processing hypothesis*, people dedicate less attention or effort to processing items when they repeat after short intervals (Greene, 1989; Hintzman, 1974). This could happen voluntarily through attentional strategies, or involuntarily through a mechanism like habituation (Smolen, Zhang, & Byrne, 2016). Inserting space between repetitions improves performance by allowing each repetition to be encoded and processed. According to the *context variability hypothesis*, items are encoded along with contextual elements present at study (Estes, 1955; Glenberg, 1979; Melton, 1970). These elements later act as retrieval cues. Because context fluctuates over time, an item is encoded along with more contextual elements when practice is temporally distributed. Spacing improves performance by increasing the probability that some of the contextual elements present at study are also present at test. Lastly, according to the *study-phase retrieval hypothesis*, the retrieval of past presentations of an item strengthens the original memory trace (Benjamin & Tullis, 2010; Bjork, 1994; Hintzman, 2004). The change in strength relates to the difficulty of the retrieval; the more difficult the successful retrieval, the greater the change. Spacing increases the difficulty of retrieval during study, potentiating the contribution of each repetition to retention.

These theories of the spacing effect are not mutually exclusive. For example, deficient processing may operate across shorter timescales spanning seconds and minutes, and context variability or study-phase retrieval may operate across longer timescales spanning hours and days (Braun & Rubin, 1998; Delaney, Verkoeijen & Spirgel, 2010). By this view, different mechanisms are responsible for spacing effects observed in single-session and multi-session studies. The exact timescales over which the different sets of mechanisms might operate remain underspecified, however. Relatedly, study-phase retrieval and context variability may jointly affect memory; the retrieval of an existing trace in a new context may allow more varied elements to be associated with it (Toppino & Gerbier, 2014; Verkoeijen, Rikers, & Schmidt, 2004). The relative contributions of the two mechanisms may be difficult to discern, in this case.

Additionally, verbal theories of this sort are general, but underspecified. Although they offer conceptual explanations for existing data, they lack the detailed mechanistic specifications needed to make precise, quantitative predictions. For example, for a fixed number of item repetitions separated by a long ITI or a short ITI, will a given retention interval be long enough to produce an advantage for spaced items? Verbal theories may also fail to make qualitative predictions in the case of new experimental designs, such as ones involving relearning. The second category of computational theories of the spacing effect do contain mechanistic specifications that enable precise, quantitative predictions. We turn now to a collection of competing computational theories.

1.4. Computational theories of the spacing effect

Multiple models have been put forth to explain how the temporal distribution of practice affects learning and retention (Mozer et al., 2009; Pavlik & Anderson, 2005; Raaijmakers, 2003; Walsh et al., *submitted*). Here we focus on three, which are among the most widely studied and used. These three were chosen because they posit different core mechanisms to capture the effects of spacing, and they relate to different theories of the spacing effect (Table 2). Other computational and mathematical models of the spacing effect exist, but they either make no predictions regarding relearning, or they contain core mechanisms already represented in the selected set (e.g., Benjamin & Tullis, 2010; Estes, 1955; Mozer et al., 2009). Distinguishing among these models is critical in terms of advancing a mechanistic account of how spacing impacts knowledge acquisition, retention, and relearning. Further, developing and validating these models creates opportunities for application. Computational models provide the rigor needed to bridge the gap between laboratory studies and educational practice (Anderson & Schunn, 2000; Dempster, 1988). For example, model outputs can be used to predict the effectiveness of new training regimes. Relatedly, model outputs can be used to predict when an individual will no longer be able to recall information, and to tailor the content and timing of training delivered to that individual (Khajah, Lindsey, & Mozer, 2014; Lindsey et al., 2014; Pavlik & Anderson, 2008).

Table 2. Core mechanisms in computational models of the spacing effect, and predictions about the impact of spacing on acquisition, retention, and relearning

			Impact of Spacing	
Model	Core Mechanisms	Acquisition	Retention	Relearning
P&A	Activation-dependent trace decay	Slowed	Increased	Slowed
PPE	Spacing-dependent trace decay	Slowed	Increased	Accelerated
SAM	Deficient processing Contextual variability Retrieval dependent updates	Slowed	Increased	No effect

The first model extends the activation equations from Adaptive Control of Thought-Rational (Pavlik & Anderson, 2005), and is henceforth referred to as P&A. In P&A, declarative knowledge is represented in the form of chunks. In paired associate recall—the paradigm used throughout this paper—each chunk contains the names of the two associates that comprise a pair. Chunks have continuously varying activation values that determine their accessibility from memory. The activation of a chunk increases with the frequency of its repetition, and decreases with the elapsed time since its repetition. Repetitions increase the accessibility of an item from memory, but as a diminishing function of its activation. Critically, the greater the activation of the item at the time of study, the higher the decay rates of additional traces added to memory. When practice is massed, an item's activation quickly rises. As a result, repetitions are stored with high decay rates. The higher decay rates associated with massed repetitions impair retention, and this has negative consequences for memory and performance over longer intervals between study and test.

P&A is generally consistent with the *study phase retrieval hypothesis*—the more difficult the retrieval, the greater the increase in an item's memory strength. One major departure from the study-phase retrieval hypothesis is that P&A does not require a successful retrieval attempt to receive these gains, as long as corrective feedback is provided. Another, more subtle, departure is that activation rather than retrieval difficulty determines an instance's decay rate. Because activation is inextricably linked to retrievability in P&A, however, items that are more difficult to retrieve always benefit more from practice than items that are easier to retrieve.

The second model is based on an extension of the General Performance Equation (Anderson & Schunn, 2000), and is called the Predictive Performance Equation, or PPE (Jastrzembski & Gluck, 2009; Walsh et al., *submitted*). PPE tracks the number of times that an item has been studied and the elapsed time since study occurred. Performance increases as a power function of number of repetitions, and decreases as a power function of elapsed time since those repetitions occurred. Retention is further modulated by the spacing between consecutive repetitions of an item (Jastrzembski & Gluck, 2009; Walsh et al., *submitted*). The greater the spacing, or lag between repetitions, the lower the decay rate. Over long intervals between study and test, the lower decay rates associated with spaced versus massed repetitions enhance retention. PPE is generally consistent with the *study phase retrieval hypothesis*. As in P&A, however, PPE does not require a successful retrieval attempt to receive these gains. Additionally, practice spacing, rather than retrieval difficulty, enhances the repetition's impact on long-term memory.

The third model is a generalization of Search of Associative Memory and is henceforth referred to as SAM (Raaijmakers, 2003). SAM is a probabilistic cue-dependent search theory. During storage, information is represented in memory traces, or "images", that contain item, associative, and contextual information. In paired associate recall, the images stored in memory correspond to word pairs. The cues that are used during retrieval are the presented member of the word pair and other contextual cues present at the time of testing. The search process consists of multiple retrieval attempts. During each attempt, an image is sampled, allowing partial recovery of information from it. The probability of responding correctly is based on the joint probability of sampling the correct image and recovering the corresponding associate from the image.

When an item is first studied, it enters a short-term memory store (STS) and a trace is created in a long-term store (LTS). If the item remains in the STS at the time of its second and subsequent presentations, repetitions are not re-encoded in memory. This is a literal instantiation of the *deficient processing hypothesis*—items that remain in the STS at the time of subsequent presentations are not re-encoded in memory. If the item has left the STS and is successfully retrieved from the LTS, contextual elements are added to the item's trace. These elements fluctuate randomly over time and act as retrieval cues during test. This is a literal instantiation of the *contextual variability hypothesis*—when practice is massed, items are encoded along with a less diverse set of contextual elements. The sparsity of contextual retrieval cues associated with items studied in a massed fashion impairs long-term retention.

All three models account for the basic spacing effect—massing practice accelerates acquisition, and spacing practice enhances retention—but in vastly different ways (Table 2). The models also account for the finding that retention performance varies non-monotonically with the length of the inter-trial interval (ITI). As the length of the ITI increases, final retention performance rises sharply before peaking and then gradually falls (Cepeda et al., 2008; Glenberg, 1976; Young, 1971). In both P&A and PPE, increasing spacing reduces decay rates, but also increases the elapsed time between earlier repetitions and the final test (Pavlik & Anderson, 2005; Walsh et al., *in revision*). If the ITI is too long, the negative effects of elapsed time may offset the benefits of reduced decay rates. The ITI that optimizes this tradeoff varies with the length of the RI (Walsh et al., *in revision*). In SAM, contextual elements are only added to a trace if the trace is successfully retrieved (Raaijmakers, 2003). This is a literal instantiation of the *study phase retrieval hypothesis*, and has been called the "retrieval-dependent update assumption" (Mozer et al., 2009). Increasing the ITI allows a more varied set of contextual elements to be added to a trace. If the ITI is too long and a retrieval failure occurs, however, new contextual elements are not added to the existing trace. The retrieval-dependent update assumption is needed for SAM to capture the non-monotonic effect of ITI on retention (Raaijmakers, 2003).

These models also make distinct, and as-of-yet untested predictions about how initial spacing will affect subsequent relearning (Table 2). P&A predicts that initial spacing may *hinder* relearning. If spaced items have higher activation than massed items at the start of the retention test, repetitions of those items during the retention test will be stored with higher decay rates (i.e., activation-dependent decay). Paradoxically, by increasing ease of retrieval, spacing reduces the benefits of relearning. SAM predicts that initial spacing may *not affect* relearning. Following retrieval failures after long retention intervals, new traces will be added to memory (i.e., retrieval dependent updates). Because these traces do not reflect initial study conditions, massed and spaced items will be relearned in an identical manner. Finally, PPE predicts that initial spacing may *accelerate* relearning. After long retention intervals, the elapsed time since an item previously appeared will be substantial, causing retrieval failures. When the item subsequently repeats, the elapsed time since it previously appeared will become far shorter. This will allow the effects of initial study conditions, still represented in the item's decay rate, to reemerge.

1.5 Overview

How does initial spacing affects subsequent relearning? This question is theoretically significant as it has the potential to distinguish between three mechanistic accounts of the spacing effect. This topic is also educationally relevant as it makes contact with the goal of enhancing ease of relearning when errorless memory is not a realistic expectation.

In the following section of the paper, we provide complete computational details of the three models, and we rigorously evaluate their predictions about the impact of spacing on learning, retention, and relearning. To foreshadow the outcome, the results of the simulation study confirm that the models make nearly identical predictions for learning and retention, but PPE uniquely predicts that initial spacing will accelerate subsequent relearning after long retention intervals. Readers primarily interested in the experimental results can bypass this section of the paper without loss of understanding.

We then report the results of two experiments motivated by the simulation study. To examine the effect of spacing on relearning, we taught participants Japanese-English word pairs (Pavlik & Anderson, 2005) in Experiment 1. Some pairs repeated after a short ITI, and some repeated after a long ITI. We tested participants' memory after retention intervals (RI) ranging from one day to three weeks. Retention tests consisted of three cued recall trials, each with corrective feedback. In this way, we could assess retention and relearning following long RIs. Based on the spacing effect literature, we expected greater retention for items studied at a long ITI. Because no spacing effect study has focused on relearning, however, it was unclear whether initial spacing would accelerate the relearning of forgotten items following long RI, as PPE predicts, or no acceleration would be found, as P& and SAM predict. Experiment 2 was a replication of Experiment 1 using a more diverse population.

2. Model Simulations Study

P&A, PPE, and SAM go beyond descriptive theories of the spacing effect in that they are instantiated as a set of mathematical equations, implemented in running computational software. These models make testable predictions, which render them falsifiable. We conducted a simulation study in order to enumerate the models' predictions regarding the impact of initial spacing on learning, retention, and relearning.

2.1 Computational implementations

The core mechanisms from each computational theory of the spacing effect are instantiated through a set of mathematical equations. We describe key equations from each of the three models in turn.[1]

2.1.1. Pavlik & Anderson (P&A). In P&A, the activation of an item depends on the frequency and recency of its presentation. These dynamics are captured by the base-level activation equation, taken from ACT-R,

$$B_n = log\left(\sum_{i=1}^{n} t_i^{-d}\right) \qquad (1)$$

where *log* is the natural logarithm of base *e*, B_n is the activation of an item that is repeated *n* times, t_i is the elapsed time between the *i*-th repetition of the item and the present time, and *d* is the decay rate. The contribution of each repetition to the item's total activation decreases with the elapsed time since it occurred (t_i^{-d}), approximating a power-law of forgetting.[2] Activation is summed across previous practice instances ($i = 1{:}n$) and logarithmically transformed, approximating a power-law of learning. The probability of successfully retrieving an item is a logistic function of its activation with slope and intercept parameters s and τ (Supplementary Material).

To account for the spacing effect, Pavlik and Anderson (2005) allowed decay (*d* in Eq. 1) to vary with an item's activation at the time of its presentation. Following a cued recall trial, the item's activation is computed (Eq. 1). A new instance of the item is added to declarative memory with an instance-specific decay rate, d_i. The decay rate is a function of the item's activation when the retrieval was attempted,

$$d_i = b + m \cdot exp(B_{n-1}), \qquad (2)$$

where *b* is a decay intercept parameter, and *m* is a decay scalar parameter. Activation is calculated with instance-specific decay rates (d_i) in the place of constant decay (cf. Eq. 1).

P&A accounts for the spacing effect in the following way. Massing practice reduces the elapsed time since earlier repetitions of an item (t_i in Eq. 1). This minimizes forgetting during initial study,

[1] All models were implemented computationally in MATLAB, and simulation code will be made available upon request.

[2] Pavlik and Anderson (2005) proposed scaling time between sessions using a constant, *h*, equal to 0.025. This effectively slows the rate of decay between sessions by reducing elapsed time. We omitted the psychological scaling of time from reported simulations and model fits in keeping with the standard implementation of time in ACT-R, and because the inclusion of scaling actually worsened the model's fit to data from the current experiments.

allowing activation to quickly rise and producing faster learning of massed items. Subsequent repetitions are stored with large decay rates, however, because of the item's high activation during study (Eq. 2). The activation of items practiced in a spaced fashion increase more gradually, slowing learning. Critically, subsequent repetitions are stored with small decay rates because of the item's lower activation during study. Across longer temporal intervals between study and test, the smaller decay rates associated with spaced repetitions enhance retention.

2.1.2. Predictive Performance Equation (PPE). In PPE, the effects of practice and elapsed time on activation (M_n) are multiplicative,

$$M_n = N^c \cdot T^{-d}. \quad (3)$$

The probability of correctly retrieving an item is a logistic function of its activation with slope and intercept parameters s and τ (Supplementary Material).

The variable N is the number of practice repetitions, T is the elapsed time, c is the learning rate, and d is the decay rate. The elapsed time T is calculated as the weighted sum of the time since each of the item's previous repetitions,

$$T = \sum_{i=1}^{n-1} w_i \cdot t_i. \quad (4)$$

The weight assigned to each repetition decreases exponentially with time,

$$w_i = t_i^{-x} \sum_{j=1}^{n-1} \frac{1}{t_j^{-x}}. \quad (5)$$

The variable x controls the steepness of weighting.[3] One could weight times equally and calculate the average elapsed time since all prior presentations of an item. However, this ignores the fact that distant events diminish in importance as new events occur (Anderson & Lebiere, 1998). Exponential weighting causes the model's representation of time to be shifted toward more recent experiences. As a consequence, although the times since each of an item's previous repetitions (t_i) necessarily increase with every subsequent repetition, elapsed time (T) may decrease because of the greater weight placed on the most recent experiences.

The variable d in Eq. 3 accounts for decay. Decay is calculated based on the history of lags, or elapsed time, between successive repetitions of an item[4],

$$d_n = b + m \cdot \left(\frac{1}{n-1} \cdot \sum_{j=1}^{n-1} \frac{1}{log(lag_j + e)} \right). \quad (6)$$

Spacing makes knowledge more resistant to decay. The quantity inside the outer parenthesis approaches zero when lags are long, reducing decay.[5] The quantity inside the outer parenthesis approaches one when

[3] In these and other reported simulations, learning rate (c) is fixed to 0.1 and x is fixed to 0.6. Previously, we found that these parameter values captured performance across 10 different published spacing effect studies (Walsh et al., *submitted*).

[4] Factors besides elapsed time, like nature of intervening activities, may impact the amount of decay that occurs between two repetitions. These factors are not currently represented in any of the three models.

[5] Lag prior to the first presentation of an item (lag_1) is ∞. Consequently, decay after the first presentation of an item is b (Eq. 6).

lags are short, increasing decay. The effects of training history are scaled by a decay slope parameter (*m*), and offset by a decay intercept parameter (*b*). Euler's number (*e* = 2.7818…) is included in the inner parenthesis to ensure that the denominator is always greater than or equal to one.

PPE accounts for the spacing effect in the following way. Massing repetitions reduces elapsed time during the acquisition phase (t_i in Eq. 4), minimizing forgetting. This allows activation to quickly rise, producing faster acquisition of massed items. The longer lags between spaced repetitions (lag_j in Eq. 6) reduce the decay rate, however. Across long intervals between study and test, the lower decay rates associated with spaced versus massed repetitions enhance retention.

PPE is based on the General Performance Equation (Anderson & Schunn, 2000), which sought to account for how amount of study and elapsed time since study impact retention. PPE goes beyond the General Performance Equation in that it represents how the temporal distribution of practice affects retention as well. PPE is also inspired by the new theory of disuse, a point that we return to in the general discussion (Bjork & Bjork, 1992).[6]

2.1.3. Search of Associative Memory (SAM). SAM is a hybrid model of the spacing effect. When a new item is presented, it enters the STS with probability *w* and a new trace is formed in the LTS. Once the item enters the STS, the probability that it remains in the buffer after *t* seconds decays exponentially (Supplementary Material). If the item remains in the STS at the time that it is next presented, the participant will respond correctly and no additional information will be added to the trace in the LTS.

Alternatively, if the item leaves STS before it is next presented, the participant must retrieve the item from the LTS to respond correctly. In each trial, SAM makes a discrete number of retrieval attempts set by the L_{max} parameter. During each attempt, an image is sampled from the LTS and information from the image is recovered. The probability of retrieving the correct associate in L_{max} attempts is

$$P_N(t) = \left[1 - \left(1 - P_{sample}(t)\right)^{L_{max}}\right] \cdot P_{recover}(t) \qquad (7)$$

The probability of sampling the correct image (P_{sample}) and the probability of recovering the missing associate given that its image is sampled ($P_{recover}$) depend on the strength of the association between the image and the current context (Supplementary Material). Contextual strength is proportional to the number of stored contextual elements active after an interval of *t* seconds that were also active during the study trial. These elements fluctuate randomly over time. The greater the time between two events, the smaller the number of contextual elements active during both. The number of elements active *t* seconds after one trial is:

$$A(t) = A(0) \cdot e^{-\alpha \cdot t} + K \cdot s \cdot (1 - e^{-\alpha \cdot t}) \qquad (8)$$

where $A(0)$ is the number of elements active at $t = 0$ and *K* is the total number of contextual elements stored in the trace. The *s* and *α* parameters control the rate of context fluctuation.

The number of elements active t_2 seconds after two trials with a spacing interval of t_1 is:

$$A_n(t_1, t_2) = A(t_1, 0) \cdot e^{-\alpha \cdot t_2} + K_2(t_1) \cdot s \cdot (1 - e^{-\alpha \cdot t_2}) \qquad (9)$$

The term $K_2(t)$ equals the total number of contextual elements stored after two trials with a spacing of t_1,

[6] A subtle distinction between P&A (and SAM) and PPE is that P&A is meant as a process model of memory (a step-by-step implementation of how memory works), whereas PPE is not. PPE is an algebraic model that generates meaningful observable and non-observable psychological variables, such as retrieval success and memory strength, but PPE's calculations are not literally implemented in the brain. Notwithstanding this consideration, PPE (along with P&A and SAM) can be used to understand how a multitude of factors combined to determine past performance, and PPE can be used to predict future performance.

$$K_2(t) = 2 \cdot A(0) - A(t_1) \quad (10)$$

That is, the total number of elements active at presentations 1 and 2 minus the number of elements active at both presentations 1 and 2 (i.e., the contextual overlap). As time between presentations increases, the contextual overlap denoted by $A(t_1)$ in Eq. 10 decreases and more contextual elements are added to the trace.

Contextual elements are added to a trace when it is created and retrieved (Supplementary Materials). Based on Eq. 8, contextual strength after one presentation is proportional to $A(t_1)$. When an item is successfully retrieved from the LTS, contextual elements are added to the trace. Based on Eq. 9, contextual strength after a *successful retrieval* is proportional to $A(t_1,t_2)$. Finally, when an item is not successfully retrieved from the LTS, a new trace is formed. Based on Eq. 9, contextual strength after an *unsuccessful retrieval* is proportional to $A(t_2)$.

SAM accounts for the spacing effect in the following two ways. First, massing practice increases the probability that items remain in the STS across successive repetitions. This increases response accuracy during initial acquisition, but effectively decreases the amount of practice that massed items receive because they are not re-encoded in memory. Second, little contextual fluctuation occurs between repetitions of massed items. The high degree of contextual overlap facilitates retrieval during initial acquisition, but causes massed items to be stored with fewer contextual elements. In contrast, spaced items are more likely to be re-encoded during initial acquisition, and to be stored with more contextual elements. Together, these factors accelerate initial acquisition for massed items, and they increase retention for spaced items.

2.2. Model Simulations

The three models all account for several major phenomena related to the temporal distribution of practice: (1) Massing accelerates acquisition; (2) Spacing increases retention; and (3) Retention varies non-monotonically with the length of the ITI. Yet the models are derived from different theoretical principles, which led to unique mechanistic accounts. The goal of this research was to distinguish between the mechanistic accounts, and in turn, their theoretical principles. Our approach was two-fold. We performed simulation studies to identify differences in model behavior as a function of spacing. We then tested empirically the predictions from these simulations.

Of the three models, P&A and PPE are most conceptually similar. In both, spaced practice reduces decay. Yet their computational implementations differ considerably. P&A maintains separate decay rates and elapsed times for each instance of an item, and activation is summed across all instances at retrieval. Subsequent repetitions do not interact with stored instances, and their contribution to activation is additive. In contrast, PPE computes the average decay rate and elapsed time for all instances. The latter quantity, elapsed time, is weighted toward more recent experiences. The positive effects of practice and the negative effects of decay are multiplied to predict performance. By reducing the average decay rate or elapsed time, repetitions can have a superadditive effect on activation.

Although the models account for the effects of spacing in similar ways, PPE's implementation leads to a unique and untested prediction: ***the temporal distribution of practice during initial study will accelerate future relearning*** (Table 2). Over long intervals between study and test, the large value of elapsed time (*T* in Eq. 3) will cause PPE's performance to asymptote near zero, producing low retention. If an item repeats during the retention test, however, the value of *T* will decrease, reflecting contributions of the item's repetition from within the session. The benefits of practice and decay (*N* and *d* in Eq. 3) will reemerge once the value of *T* no longer causes PPE's performance to saturate near zero. In other words, the effects of initial practice conditions (e.g., number of repetitions or spacing) will persist in PPE, although they will only reappear once knowledge is reactivated by relearning.

P&A predicts that spacing will improve retention but may limit gains from relearning trials. Decay rates for item repetitions increase with the item's activation at the time of test (Eq. 3). If the benefits of spacing remain at the start of the retention test, repetitions of spaced items will be stored with higher decay rates. In other words, although spacing may improve retention in P&A, it will limit gains from relearning trials.

SAM is quite different from P&A and PPE, yet its predictions about relearning are consistent with P&A. SAM predicts that the effects of initial learning conditions will be lost if items cannot be retrieved at the start of the retention test. This is a direct consequence of the retrieval-dependent update assumption. If an item is not successfully retrieved, a new trace is added to the LTS (Raaijmakers, 2003). The effects of initial study conditions, including the set of contextual elements encoded in the original trace, are lost once the new trace is added to the LTS.

To make the models' predictions and their differences more concrete, we simulated an experiment in which participants learned and were repeatedly tested on paired associates during an acquisition phase. They relearned and were repeatedly tested on the same paired associates during a retention phase. In each trial, a cue was presented and the model attempted to retrieve its associate. The model then received feedback about the correct associate. The acquisition phase consisted of 15 test-restudy trials. Half of the associates repeated after a short inter-trial interval (ITI) averaging 10 seconds, and half repeated after a long ITI averaging 60 seconds. The retention phase was administered 1 or 21 days later, and consisted of 3 additional test-restudy trials. During the retention phase, all items repeated after the same ITI averaging 60 seconds.

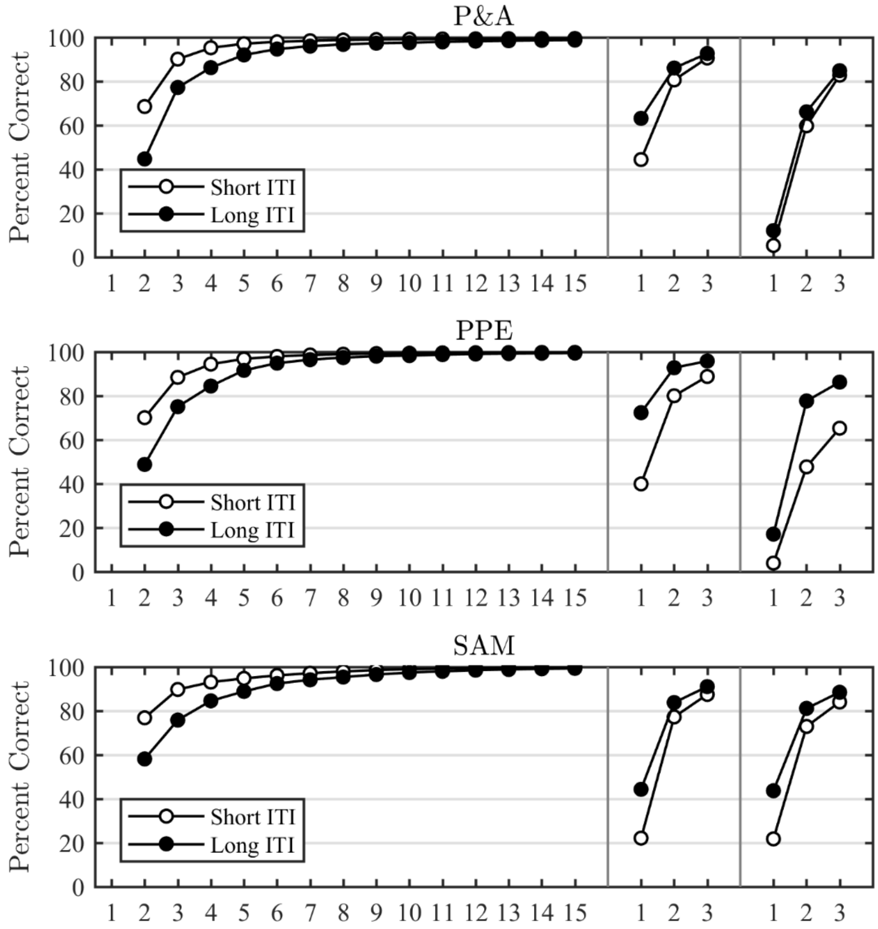

Figure 1. Predictions from P&A (top), PPPE (middle), and SAM (bottom) for simulated experiment. Trial numbers are marked along the x axis.

Figure 1 shows predictions of the three models for the Acquisition Phase (left) and for the 1 and 21 Day Retention Tests (right). For the Acquisition Phase, all models predict the gradual mastery of items, and all predict that items presented at a short ITI (open circles) will be acquired faster than items presented at a long ITI (filled circles). On the 1 Day Retention Test, all models predict that items presented at a long ITI will be retained better than items presented at a short ITI. Together, predictions from the Acquisition Phase and the 1 Day Retention Test are consistent with the standard spacing effect. On the 21 Day Retention Test, memory performance begins near zero. Critically, although all models predict rapid relearning, PPE uniquely predicts that the effects of the spacing manipulation, administered only during the acquisition phase, will reemerge during the second and third trials of the retention test. Both P&A and SAM predict that if the effect of the initial spacing manipulation is no longer present at the start of the retention test, all items will be relearned at the same rate.

The key feature of Figure 1 is how the performance differences between long and short ITI items change across the three trials of the 21 Day Retention Test. The difference increases across relearning trials in PPE, reflecting spacing-accelerated relearning, whereas it remains the same or decreases across relearning trials in P&A and SAM. Although the predictions in the graphs are based on one parameterization of each model, they are illustrative of the qualitative trends that nearly all parameterizations of each model produce. We ran P&A and PPE using 160,000 parameterizations and SAM using 1,000,000 parameterizations (parameter details are described in Section 4.1), and calculated the difference in performance between long and short ITI items during the 21 Day Retention Test. The mean differences on trials 1 and 3 of the retention test were 9.7% and 12.5% for PPE, 5.8% and 1.2% for P&A, and 5.6% and 0.5% for SAM. In other words, PPE consistently predicts that the effect of initial spacing will increase across item repetitions during the retention phase, whereas P&A and SAM predict that the effect will decrease.

2.3. Discussion

The *qualitative* differences in model predictions provide an opportunity to distinguish between PPE versus P&A and SAM experimentally. The simulations showed that PPE is incapable of mimicking P&A and SAM, and vice versa, so the results of an empirical test could be similarly definitive. If the results show that initial spacing does not affect relearning, thereby rejecting PPE, the *quantitative* differences in model predictions provide a further opportunity to distinguish between P&A and SAM. In particular, Pavlik and Anderson (2005) noted that SAM failed to account for the wide range of performance values across trials and conditions observed in their experiment. This is partially evident in Figure 1, which shows that SAM predicts a narrower range of performance values than either PPE or P&A. SAM predicts higher levels of performance at the start of the Acquisition Phase, and higher levels of retention after long (i.e., greater than 1 day) retention intervals. Given this difference, the results of an empirical test are also likely to be reasonably definitive in distinguishing between P&A and SAM.

3. Experiment 1

The simulation study revealed that PPE predicted that initial spacing would accelerate subsequent relearning, even when retention was extremely low. P&A and SAM, in contrast, predicted no effect—or a negative effect—of initial spacing on subsequent relearning. We conducted two experiments with human participants to evaluate the models' predictions.

3.1. Material and Methods

3.1.1. Participants

Twenty-two individuals from The Ohio State University participated in a multi-week study for monetary compensation (8 females and 14 males, ages ranging from 18 to 28 years with a mean age of 21 years). They received $100 for completing five experiment sessions. Two additional participants failed to complete all sessions, and data from a third participant were excluded due to an equipment error. The study was approved by the Institutional Review Board at Ohio State University.

3.1.2. Task

Participants memorized Japanese-English word pairs based on materials used in a study by Pavlik and Anderson (2005). When a pair was presented for the first time, the Japanese word and the English translation appeared simultaneously on the computer screen. Participants typed the translation using a standard keyboard. During each subsequent presentation, the Japanese word appeared alone (Figure 2). Participants were instructed to recall and type the translation. After they responded or after 7 seconds had passed, positive or negative feedback in the form of a smiling or frowning face appeared, and the correct response was given. Feedback remained on the screen for 2 seconds, and trials were separated by a blank inter-trial interval of 0.5 seconds. The task program was written and executed in MATLAB using the Psychophysics Toolbox extensions (Brainard, 1997), and was administered on standard desktop computers in a laboratory in The Ohio State University's Psychology Department.

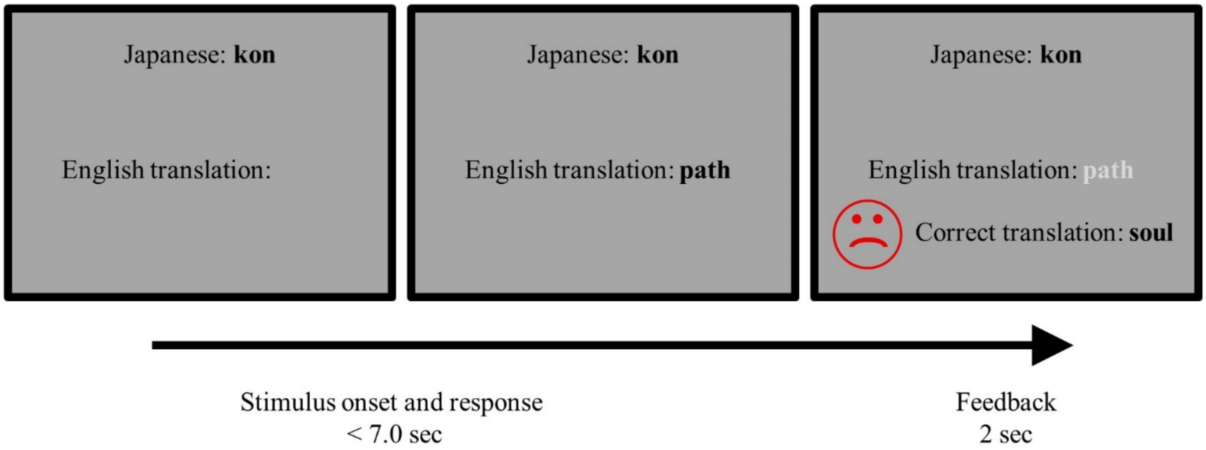
Japanese: **kon**
English translation:
Japanese: **kon**
English translation: **path**
Japanese: **kon**
English translation: **path**
Correct translation: **soul**
Stimulus onset and response
< 7.0 sec
Feedback
2 sec

Figure 2. Experiment task. After a 0.5 s inter-trial interval, a Japanese word appeared. Feedback along with the correct response appeared once the participant responded or after 7 s passed.

Japanese words contained from 3 to 7 letters and had a mean length of 5.5 letters. English translations all contained 4 letters. Words were selected from the MRC Psycholinguistic database and had familiarity ratings from 430 to 621 with a mean of 545, and imaginability ratings from 362 to 592 with a mean of 482. The words were chosen to have higher familiarity and imaginability ratings than the overall MRC database means of 495 and 394, respectively.

3.1.3. Procedure and Schedule

Participants completed 5 sessions over the course of 44 days (Table 3). During Session 1, they learned a set of Japanese-English word pairs. During Sessions 2, 3, and 4 they were tested on word pairs from the previous session and they learned new sets of word pairs. During Session 5, they were tested on word pairs from the previous session and they were then tested on all word pairs from the entire experiment. We call this the *final* retention test to distinguish it from the *intermediate* retention tests administered at the start of Sessions 2 through 5.

We manipulated the retention intervals (RIs) between acquisition phases and retention tests. Each participant completed retention tests after 1, 7, 14, and 21 day RIs. We controlled RI by varying the amount of time between sessions. For example, to administer RIs in ascending order, participants were scheduled for days 1, 2, 9, 23, and 44 (Table 3). This created intervals of 1, 7, 14, and 21 days between when a set was learned and when it was tested. We created four different RI sequences according to a Latin square (1-7-14-21; 7-21-1-14; 14-1-21-7; and 21-14-7-1), and randomly assigned sequences to participants.[7] We chose intervals of 1, 7, 14, and 21 days based on the outcome of the simulations described earlier, our theoretical interest in observing retention and relearning after short, medium, and long retention intervals, and practical considerations such as identifying convenient and predictable times for participants to return to the laboratory.

Table 3. Schedule and composition of experiment sessions

Session	Day	Composition
1	1	Acquisition phase: Set A (350 trials)
2	2	Retention test: Set A (60 trials) Acquisition phase: Set B (350 trials)
3	9	Retention test: Set B (60 trials) Acquisition phase: Set C (350 trials)
4	23	Retention test: Set C (60 trials) Acquisition: Set D (350 trials)
5	44	Retention test: Set D (60 trials) Final retention test: Sets A, B, C, D (240 trials)

We also manipulated the inter-trial intervals (ITIs) between item repetitions within acquisition sessions. Items were assigned to a short ITI condition with 1 intervening trial between repetitions, or a long ITI condition with 9 intervening trials between repetitions. Participants were not told about the short and long ITIs, and items from both conditions were intermixed. During retention tests, all items repeated after 9 intervening trials; that is, the ITI manipulation was restricted to initial acquisition.

For each of the 4 RIs, participants learned and were tested on 20 distinct Japanese-English word pairs (Table 3). Half were presented at the short ITI, and half were presented at the long ITI. Each word pair repeated 15 times during the acquisition phase, 3 times during the intermediate retention test, and 3 times during the final retention test. To prevent a primacy effect (Crowder, 1976), we included filler items in each session. Participants completed 11 to 25 trials with filler items (with a mean of 17 trials) at the start of the acquisition phase in each session. Additionally, to prevent participants from being able to perfectly anticipate repetitions of short ITI items in alternating trials, we inserted unpredictable fillers throughout the acquisition phase. This increased the ITI to 1.25 trials in the short condition, and 10.25 trials in the long condition. In total, the acquisition phases had 350 trials (including 50 filler trials), the intermediate retention tests had 60 trials, and the final retention test had 240 trials. Participants were given a 60-second break after completing the retention test at the start of Sessions 2 through 5, and midway through acquisition phase in each session. During breaks, they saw their average accuracy since the previous break, and their highest accuracy from any completed test to that point in the experiment. Participants took an average of 35 minutes to complete each session.

3.2. Results

Each participant completed the acquisition phase a total of four times during Sessions 1 through 4. We averaged across sessions, and analyzed acquisition performance as a function of item repetition number within the session. Each participant also completed four intermediate retention tests after different RIs along with a final retention test. We analyzed performance separately for each retention test and as a function of item repetition number. Acquisition sessions and retention tests included 10 items presented at short ITIs and 10 presented at long ITIs. We averaged response accuracy across the items that made up a schedule, and across the 22 participants who completed the experiment.

Figure 3 shows response accuracy during the acquisition phase (top panel), and for the intermediate and final retention tests (bottom panel). We analyzed data using logistic mixed-effects

[7] All sessions were completed within 2 days of the target dates, and 88% were completed on the target dates. The mean values obtained for the RIs (1.3, 7.1, 14.1, and 20.8 days) were very close to the intended lengths (1, 7, 14, and 21 days).

regression with the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in the *R* statistical computing environment. We performed separate regression analyses on data from the acquisition phase, the intermediate retention tests, and the final retention test. In all analyses, we treated accuracy (correct or incorrect) as the response variable, ITI, and trial number as fixed factors, and item and subject as random factors. We included RI as an additional fixed factor in our analysis of data from the intermediate retention tests. Results from all logistic mixed-effects regression analyses are contained in Supplementary Table 1.

We analyzed data from repetitions 2 through 15 during the acquisition phase (the full word pair appeared during the first presentation, and so the first trial did not provide a meaningful performance measure). Accuracy increased with repetition and was higher for items presented at a short ITI, reflecting main effects of repetition ($\beta = 1.28$, $SE = 0.05$, $z = 23.75$, $p < .001$) and ITI ($\beta = -0.56$, $SE = 0.09$, $z = -5.97$, $p < .001$). This replicates the typical advantage for massed practice during acquisition. The interaction between ITI and trial was also significant ($\beta = 0.38$, $SE = 0.54$, $z = 7.12$, $p < .001$). Although accuracy was initially higher for items presented at a short ITI, participants mastered all items by the end of the acquisition phase.

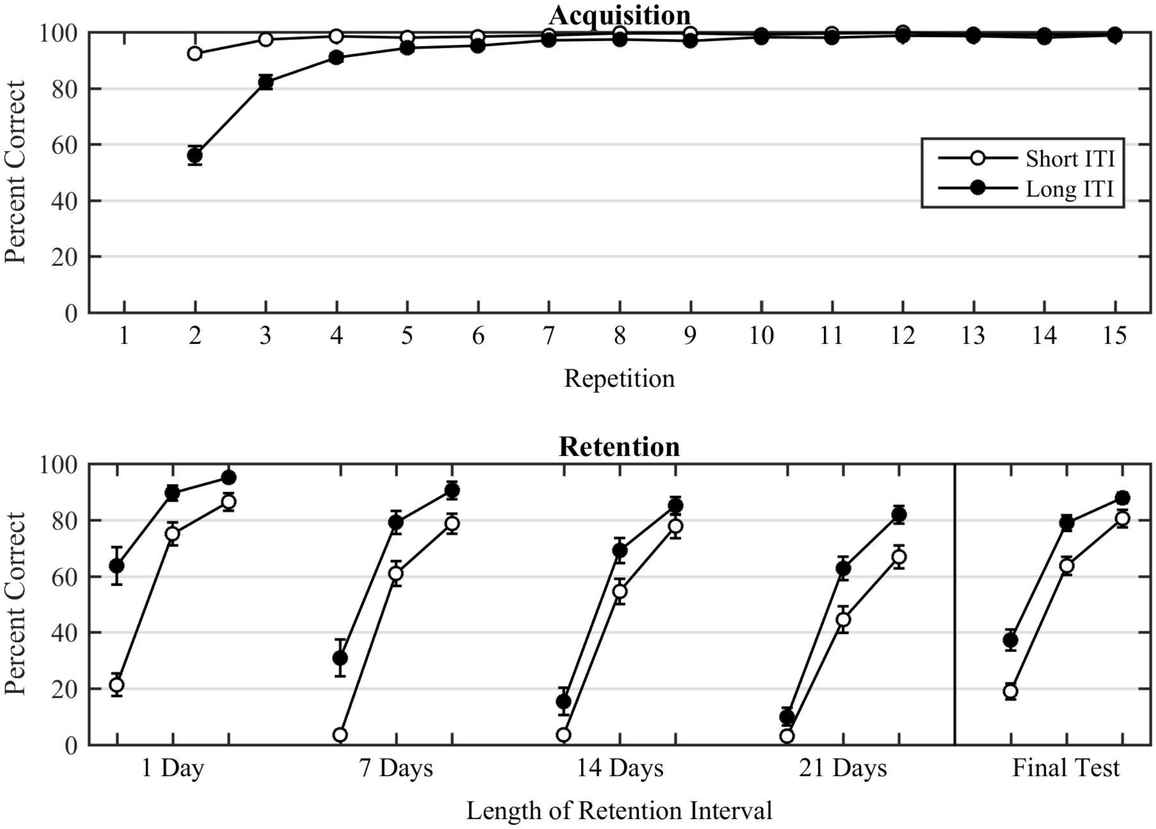

Figure 3. Performance during the acquisition phase (top), and the intermediate and final retention tests (bottom). Error bars show ±1 standard error of the mean.

Next, we analyzed performance during the four intermediate retention tests. Accuracy decreased with the length of the retention interval (1, 7, 14, and 21 days), consistent with the standard effect of forgetting ($\beta = -0.60$, $SE = 0.04$, $z = -15.51$, $p < .001$). Additionally, and in contrast to the acquisition phase, accuracy was higher for items initially studied at a long ITI ($\beta = 0.51$, $SE = 0.06$, $z = 8.37$, $p < .001$). This replicates the typical advantage for spaced practice during retention. Most interestingly, the three-way interaction between ITI, RI, and trial number was significant ($\beta = 0.09$, $SE = 0.04$, $z = 2.34$, $p < .05$). When the retention interval was short (e.g., 1 Day), the relative benefit of spaced versus massed practice was greatest on Trial 1 of the retention test. This reflects a *retention* advantage. Alternatively, when the retention interval was long (e.g., 21 Days), the relative benefit of spaced practice was almost entirely absent on the first trial of the retention test but reemerged by Trial 3. This reflects a *relearning* advantage. To probe the nature of the three-way interaction while controlling for floor and ceiling effects, we replotted results using a logit scale (Supplementary Figure 1). The interpretation of the interaction remained the same.

The relearning advantage for spaced items was actually present after every retention interval, but was obscured by the overlapping retention advantage still present after shorter intervals. To isolate relearning effects, we identified the subset of items not correctly recalled on the first trial of each retention test. We then analyzed response accuracy to those items on their second and third presentations (Figure 4). Of the items not correctly recalled on the first trial of a retention test, those initially studied at a long ITI were relearned more quickly ($\beta = 0.29$, $SE = 0.06$, $z = 5.22$, $p < .001$). Importantly, the interaction between ITI and retention interval was not significant ($\beta = 0.03$, $SE = 0.05$, $z = 0.88$, *n.s.*). After every retention interval, forgotten items that were initially presented at a long ITI were relearned more quickly.

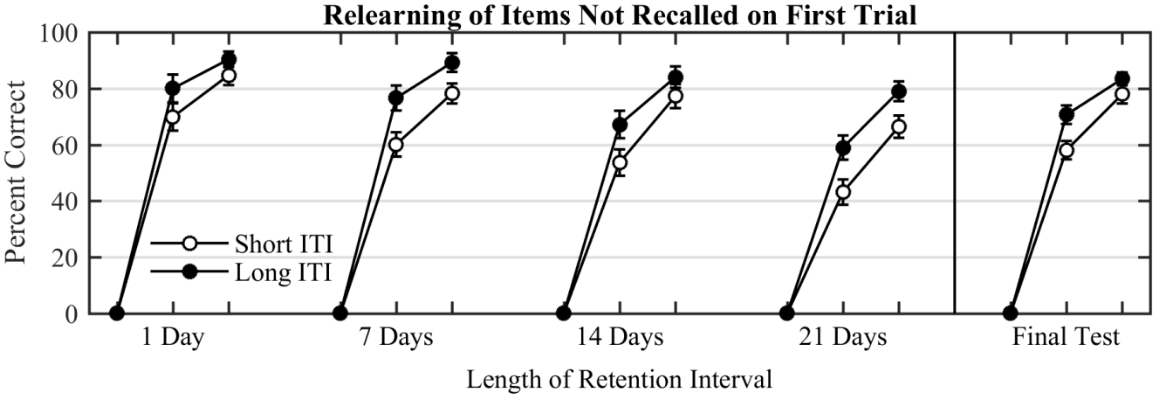

Figure 4. Performance during the final retention tests for the subset of items not recalled on the first trial (bottom). Error bars show ±1 standard error of the mean.

Lastly, we analyzed performance during the final retention test for all word pairs (Figure 3). Paralleling results from the intermediate retention tests, accuracy was higher for items presented at a long ITI ($\beta = 0.42$, $z = 5.99$, $SE = 0.07$, $p < .001$), and increased across trials ($\beta = 1.32$, $SE = 0.04$, $z = 32.09$, $p < .001$). The interaction between ITI and trial was not significant ($\beta = -0.05$, $SE = 0.04$, $z = -1.35$, *n.s.*). As with the intermediate retention tests, we identified the subset of items not correctly recalled during the first trial of the final test, and we analyzed response accuracy to those items during Trials 2 and 3 (Figure 4). Forgotten items initially presented at a long ITI were relearned more quickly ($\beta = 0.23$, $SE = 0.07$, $z = 3.65$, $p < .001$).

3.3. Experiment 1 Summary

Increasing the spacing between item repetitions slowed acquisition but improved retention, replicating the standard spacing effect (Cepeda et al., 2006). As the length of the retention interval increased from 1 to 21 days, retention decreased, replicating the standard forgetting function (Wixted & Ebbesen, 1997). Most importantly, the benefits of spacing were greatest for early trials of the retention test when the RI was short, and for later trials when the RI was long. The interaction was driven by the relearning of the subset of items not correctly recalled at the start of the retention tests. Of those items, the ones initially presented at a long ITI were relearned more quickly. The effect of initial spacing persisted during the final retention test, which is even more remarkable given that all items were presented with the same ITI during the intermediate retention tests that came between initial acquisition and the final test.

These results establish the two-fold benefits of spacing on boosting retention after short RIs and accelerating relearning after all RIs.

One potential concern with analyzing the subset of items not correctly recalled at the start of the retention test is that the analysis may be contaminated by item selection effects; in other words, items not recalled may have differed in some systematic between the short and long ITI conditions. We think such effects would actually works *against* the hypothesis of spacing-enhanced relearning. Items not recalled in the long ITI condition should be more difficult items on average than those not recalled in the short ITI condition. This makes the finding that spaced items were re-learned more quickly even more convincing. Another concern is that incorrect responses could arise for reasons besides retrieval failures, such as attentional lapses. Although attentional lapses may have occurred, any influence they had would be minor given the uniformly high error rates following long retention intervals.

4. Experiment 2

The finding that spacing accelerates relearning is entirely novel. This finding was predicted by PPE, but not by P&A or SAM. Given the theoretical importance of this finding in discriminating between the models, and more generally its relevance to educational practices, we replicated the experiment to evaluate the robustness of the effect (Maner, 2014; Simons, 2014).

We administered the same task to a population of participants recruited through Amazon Mechanical Turk (MTurk). As compared to college undergraduate and graduate students, the MTurk population is more diverse, older, and less educated (Buhrmester, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014). In addition to differences between the participant populations in Experiments 1 and 2, the task settings necessarily differed. Participants in Experiment 1 completed the study in a carefully controlled laboratory environment, whereas participants in Experiment 2 could complete the study in an environment of their choosing. We expected that administering the study to a more diverse population and in a less controlled setting would reduce the overall level of performance. Previous studies have reported lower baseline data quality in MTurk versus laboratory studies (Goodman, Cryder, & Cheema, 2013; Simcox & Fiez, 2014). However, we did not expect that this change would modulate the critical effects of spacing on retention and relearning.

4.1. Material and Methods

The task, procedure, and schedules used in Experiments 1 and 2 were identical. Participants in the United States freely registered, and they completed the experiment in a non-laboratory setting using a personal computer. The MATLAB program from Experiment 1 was translated verbatim into a combination of Javascript and PHP to run on an Apache server in the Ohio State University Psychology Department. The task was presented on a webpage that was accessed through MTurk. Each data file was marked with the participant's unique and anonymous MTurk Worker ID to collate files from multiple sessions. Target dates for completing sessions were emailed to participants.[8]

A total of 25 individuals finished the study (10 females and 15 males, ages ranging from 23 to 58 years with a mean age of 36 years). They received $42 for completing five experiment sessions. 12 additional participants failed to complete all sessions. The numbers of participants completing Sessions 1 through 5 were 37, 33, 31, 29, and 25, equating to about a 10% dropout rate per session. Participants who failed to complete all sessions performed slightly worse during Session 1 (85% versus 91%, $t(35) = 2.12$, $p < .05$), indicating a slight tendency for lower-performing individuals to drop out.

4.2. Results

Figure 5 shows response accuracy as a function of item repetition during acquisition (top panel), and for the intermediate and final retention tests (bottom panel). Data were aggregated across items within

[8] All sessions were completed within 2 days of the target dates, and 98% were completed on the target dates. The mean values obtained for the RIs (1.00, 7.01, 14.05, and 21.0 days) were extremely close to the intended durations (1, 7, 14, and 21 days).

schedules and across participants as in Experiment 1. Results from all logistic mixed-effects regression analyses are contained in Supplementary Table 2.

The results replicate those found in Experiment 1. During acquisition, response accuracy was higher for items presented at a short ITI (*β* = -0.75, *SE* = 0.06, *z* = -12.34, *p* < .001). This reversed during intermediate and final retention tests, where accuracy was higher for items presented at a long ITI, (intermediate retention tests: *β* = 0.44, *SE* = 0.04, *z* = 11.18, *p* < .001; final retention test: *β* = 0.28, *SE* = 0.06, *z* = 5.00, *p* < .001). Most importantly, the effects of ITI and RI during intermediate retention tests were modulated by trial number (*β* = 0.11, *SE* = 0.04, *z* = 3.20, *p* < .01). When the retention interval was short (i.e., 1 Day), the relative benefit of spaced versus massed practice was greatest on the first trial, and when the retention interval was long (i.e., 21 Days), the relative benefit of spaced practice was absent on the first trial but re-emerged by the third trial. To examine the three-way interaction while controlling for floor and ceiling effects, we replotted results using a logit scale (Supplementary Figure 2). The interpretation of the interaction remained the same.

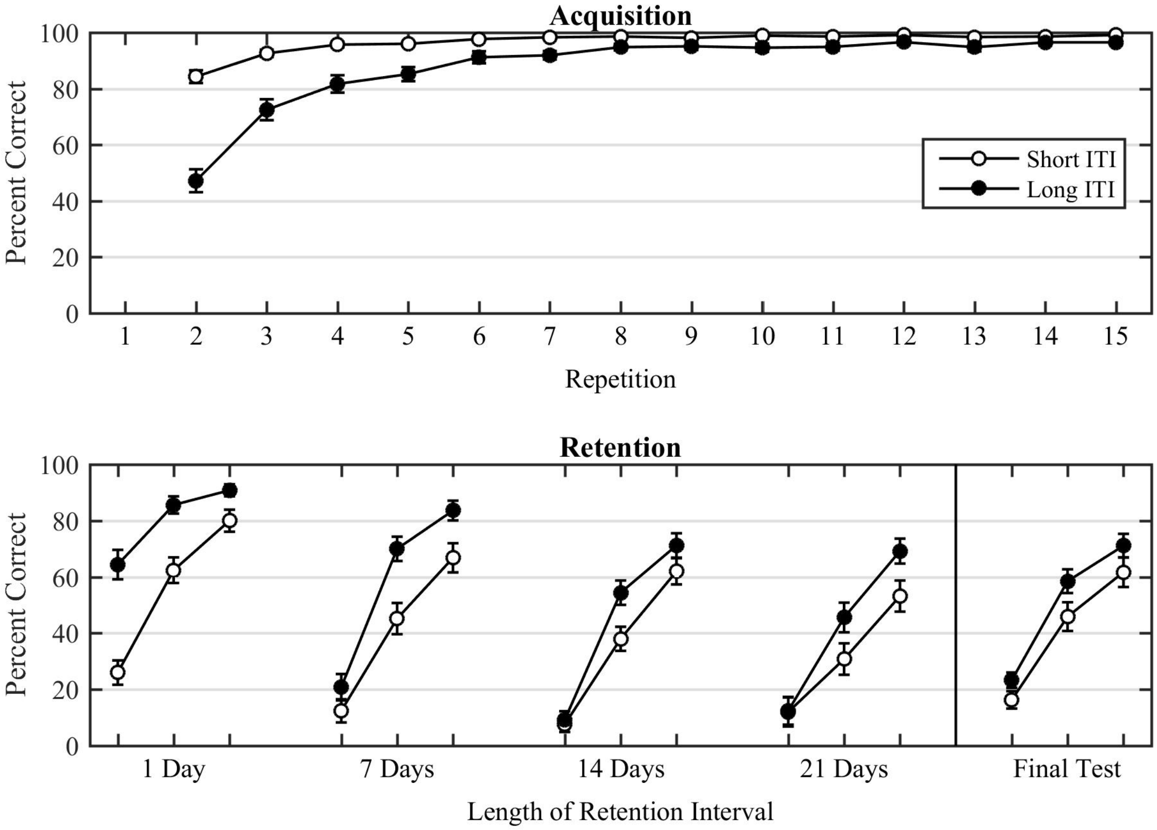

Figure 5. Performance during the acquisition phase (top), and the intermediate and final retention tests (bottom). Error bars show ±1 standard error of the mean

As in Experiment 1, the relearning advantage for spaced items was present after every retention interval, but was obscured by the retention advantage still present after shorter intervals. To isolate relearning effects, we identified the subset of items not correctly recalled on the first trial of the retention test and analyzed response accuracy for those items on their second and third presentations (Figure 6). Of the items not correctly recalled on the first trial of the retention test, those initially studied at a long ITI were relearned more quickly (*β* = 0.39, *SE* = 0.05, *z* = 8.02, *p* < .001). Importantly, the interaction between ITI and retention interval was not significant (*β* = 0.00, *SE* = 0.04, *z* = 0.07, *n.s.*). Across *all* retention intervals, forgotten items that were initially presented at a long ITI were relearned more quickly.

We applied the same analysis to items not correctly recalled on the first trial of the final retention test. Again, items initially presented at a long ITI were relearned more quickly (*β* = 0.46, *SE* = 0.10, *z* = 4.49, *p* < .001).

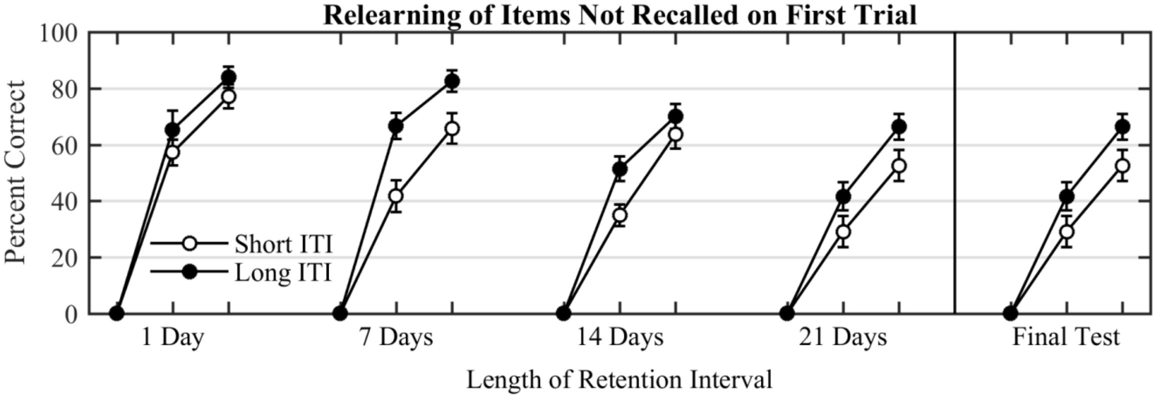

Figure 6. Performance during the final retention tests for the subset of items not recalled on the first trial (bottom). Error bars show ±1 standard error of the mean.

Although the data patterns were strikingly similar, we compared performance between Experiments 1 and 2 to assess whether the different population or testing environment altered performance. During the acquisition phase and the final retention test, there was a main effect of experiment reflecting the greater accuracy in Experiment 1 (acquisition: *β* = -0.36, *SE* = 0.10, *z* = -3.53, *p* < .001; final retention: *β* = -0.45, *z* = -3.39, *p* < .001). During the intermediate retention tests, there was a marginally significant effect of experiment (*β* = -0.21, *SE* = 0.11, *z* = -1.86, *p* < .1). Most importantly, the three-way interaction between ITI, RI, and relearning trial was significant (*β* = 0.10, *SE* = 0.03, *z* = 3.81, *p* < .001) and was not modulated by experiment (*β* = 0.01, *SE* = 0.03, *z* = 0.35, *n.s.*). In both experiments the effect of initial learning conditions faded across relearning trials following short RIs, and grew across relearning trials following long RIs.

4.3. Experiment 2 Summary

Despite differences in testing sample and environment, Experiment 2 replicated all key outcomes from Experiment 1, demonstrating the generality and the robustness of the findings. Short ITIs facilitated performance during the acquisition phase, whereas long ITIs facilitated performance during retention tests. Most importantly, the effect of ITI on performance was modulated by the length of the RI and trial number—the effect of initial spacing was greatest on the first trial of retention tests delivered after a short interval, and on the second and third trials of retention tests delivered after a long interval. This interaction arose from relearning on the subset of items not correctly recalled at the onset of the intermediate and final retention tests. Of those items, the ones initially presented at a long ITI were relearned more quickly. The relearning effect was very robust. The effect appeared in the group averaged data from all four intermediate retention tests and the final retention test in both experiments. Averaging across retention intervals, the relearning effect appeared in 20 of 22 participants in Experiment 1 and 21 of 25 in Experiment 2. Together, the results from Experiments 1 and 2 not only replicate the standard finding that spacing enhances retention, they also show for the first time, and quite consistently, that spacing accelerates relearning.

Before assessing how well the models account for the data quantitatively, we address a few concerns about the interpretation of the results. One is what to make of the ceiling and floor effects present in Figures 3 and 5. Are our conclusions hampered by these effects? We think not. Performance data from retention tests delivered after shorter intervals (i.e., 1 and 7 days) do contain a ceiling effect. Because performance on Trial 1 is higher for long ITI items, and because response accuracy is bounded by 100%, the relearning curve for long ITI items is necessarily shallower than the curve for short ITI items. This simple fact illustrates the challenge of detecting a relearning advantage separately from the retention advantage. However, restricting analysis to the subset of items not correctly recalled at the start of the retention tests eliminated the ceiling effect (Figures 4 and 6). This analysis revealed that long ITI items were relearned more quickly after every retention interval.

Performance data from the retention tests delivered after longer intervals (i.e., 14 and 21 days) do contain a floor effect. The spacing effect may persist, but may be too small to detect at the start of the retention tests given after such long intervals. In fact, this is one justification for administering multiple relearning trials rather than one test trial—the speed of relearning may reveal benefits of initial learning conditions that are not evident based on retention alone (de Jonge et al., 2014; MacLeod, 1976; 1988; Mazza et al., 2016). This methodological detail aside, the re-emergence of a spacing effect across relearning trials for forgotten items is highly diagnostic with respect to computational models and theories of the spacing effect, as we describe in the coming sections.

A final, related concern is that spacing enhances memory, and that any item with enhanced memory may be relearned more quickly. According to this argument, *any* manipulation that enhanced memory would accelerate relearning. Further, differences in memory strength that are too weak to detect using recall could nonetheless accelerate relearning. To address this concern, one would need to demonstrate spacing-accelerated relearning when initial retention performance is above floor and initial retention of spaced items is less than or equal to retention of massed items. Unfortunately, if all massed and spaced items are presented the same number of times during initial acquisition and if all are tested after the same retention interval, these conditions will not likely be met. If the retention interval is short (i.e., 1 and 7 days), performance will be above floor but the retention advantage for spaced items will remain. Alternatively, if the retention interval is long (i.e., 14 and 21 days), the retention advantage for spaced items will be diminished but performance will be at floor.

Nevertheless, to address these issues, we performed an additional sub analysis. We compared retention of spaced items after a long interval against retention of massed items after a short interval.[9] In such a comparison, the length of the interval offsets the benefits of spacing on retention. This has the effects of keeping performance from the massed condition above the floor at Trial 1, and of keeping performance from the spaced condition below ceiling at Trial 3. The data from Experiment 2 enables this check – initial retention after 1 day in the short ITI condition exceeded initial retention after 7 days in the long ITI condition, and initial retention after 7 days in the short ITI condition exceeds initial retention after 14 days in the long ITI condition. These data are contained in Figure 3 and 5, and are replotted for clarity in Figure 7.[10] We performed a logistic mixed-effects regression with RI (short and long), ITI, and relearning trial as fixed factors, and item and subject as random factors. Memory performance decreased with the length of retention interval ($\beta = -0.42$, $SE = 0.05$, $z = -9.13$, $p < .001$), increased across relearning trials ($\beta = 1.31$, $z = -24.59$, $p < .001$), and did not vary with spacing ($\beta = 0.09$, $SE = 0.05$, $z = 1.58$, *n.s.*). Most importantly, the interaction between spacing and trial was significant ($\beta = 0.12$, $SE = 0.05$, $z = 2.37$, $p < .05$), reflecting the fact that accuracy on Trial 1 was highest for short ITI items, whereas accuracy on Trials 2 and 3 was highest for long ITI items. This confirms that the relearning advantage for spaced items is present even in a subset of the data that avoids floor and ceiling effects, and that equates performance between conditions on the first trial of the retention test

[9] Length of retention interval is not confounded with task experience because the order of retention intervals was counterbalanced across participants. Otherwise, items encountered after longer retention intervals might benefit from general task learning effects.

[10] The corresponding analysis from Experiment 1 is less informative because accuracy for long ITI items following 7 and 14 days exceeded accuracy for short ITI items following 1 and 7 days.

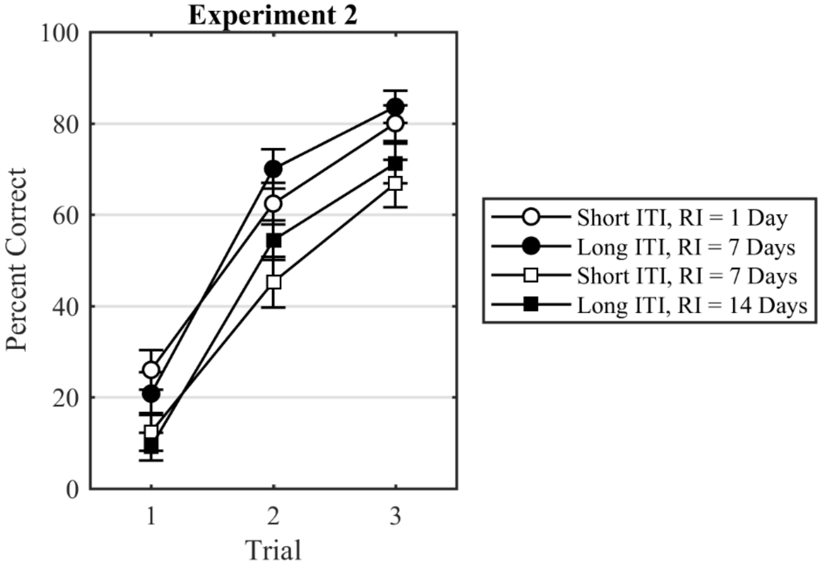

Figure 7. Performance during three trials of retention test based on whether items were initially studied at short or long ITIs, and whether the retention interval was 1, 7, or 14 days.

5. Model Comparison

The experiments were motivated by three computational models of the spacing effect. One model predicted that spacing would accelerate relearning and the other two did not. Qualitatively, the results from the intermediate retention tests were most consistent with PPE – participants displayed a relearning advantage following spaced practice. This was only one of the experimental outcomes, however, making it unclear which of the three models could best account for the complete set of results involving the effects of spacing on learning, retention, and relearning. For example, the parameter values in PPE needed to produce spacing-enhanced relearning might over-predict or under-predict the effects of spacing on acquisition or retention. A novel feature of our design was that all experimental factors (ITI and RI) were crossed and administered within-subject. A major advantage of this design is the greater constraint it provides in terms of fitting computational models to individual participants' data. The requirement to simultaneously account for the effects of trial-level spacing on acquisition, retention, and relearning across intervals ranging from 1 day to 3 weeks for individual participants provides an extremely strong test of the three models.

Our evaluation of the models was guided by three questions:

1. Can the models simultaneously capture the effects of spacing on learning, retention, and relearning?
2. Do the results from individual participants consistently favor one model?
3. When withholding data from subsets of the retention tests, can the models predict performance from the omitted tests?

The first and second questions concern the theoretical adequacy of the models in accounting for the empirical phenomena. The third question, in addition to providing an additional test of theoretical adequacy, also speaks to the application potential of the models, such as in an educational setting. If a model accurately predicts future performance, the model can be used to prescribe when to administer refresher training and review. All details on parameter estimation and modeling fitting are contained in the supplementary material.

5.1. Results

5.1.1. Model Fits

Do the models simultaneously capture effects of learning, retention, and relearning? To address this question, we estimated the model parameter values that maximized the log-likelihood of the data from each participant (Myung, 2003). We then aggregated model fits across items within a schedule and across participants exactly as with the experimental data. Figure 8 shows the fit of P&A to the complete data from Experiment 1. The results from Experiment 2, which were comparable, are shown in the Supplementary Materials.

The experimental data contained four key effects, three of which P&A accounted for. First, during the acquisition phase, participants learned items presented at short ITIs more quickly. P&A acquired items presented at a short ITI more quickly because elapsed time since earlier repetitions (*T* in Eq. 1) was less for those items than for items presented at a long ITI. Second, accuracy at the start of intermediate tests decreased with the length of the retention interval. Retention in P&A decreased with the length of the retention interval owing to the greater elapsed time (*T* in Eq. 1), and hence greater decay, from study to test. Third, accuracy during retention tests was higher for items initially studied at long ITIs. Retention in P&A was greater for items initially studied in a spaced manner. When an item repeated after long ITIs during the acquisition phase, its activation increased more gradually – this is reflected in the lower accuracy for spaced versus massed items during the acquisition phase. Because activation was lower for spaced items, repetitions were stored with lower decay rates (*d* in Eq. 2). Across the longer intervals between acquisition and test, the lower decay rates associated with spaced repetitions improved retention. The relative benefits of spacing on retention where substantially larger in the experimental data than in P&A, however. Fourth and finally, spacing accelerated relearning. This was most apparent after long retention intervals. P&A failed to produce a relearning advantage for items presented at long ITIs.

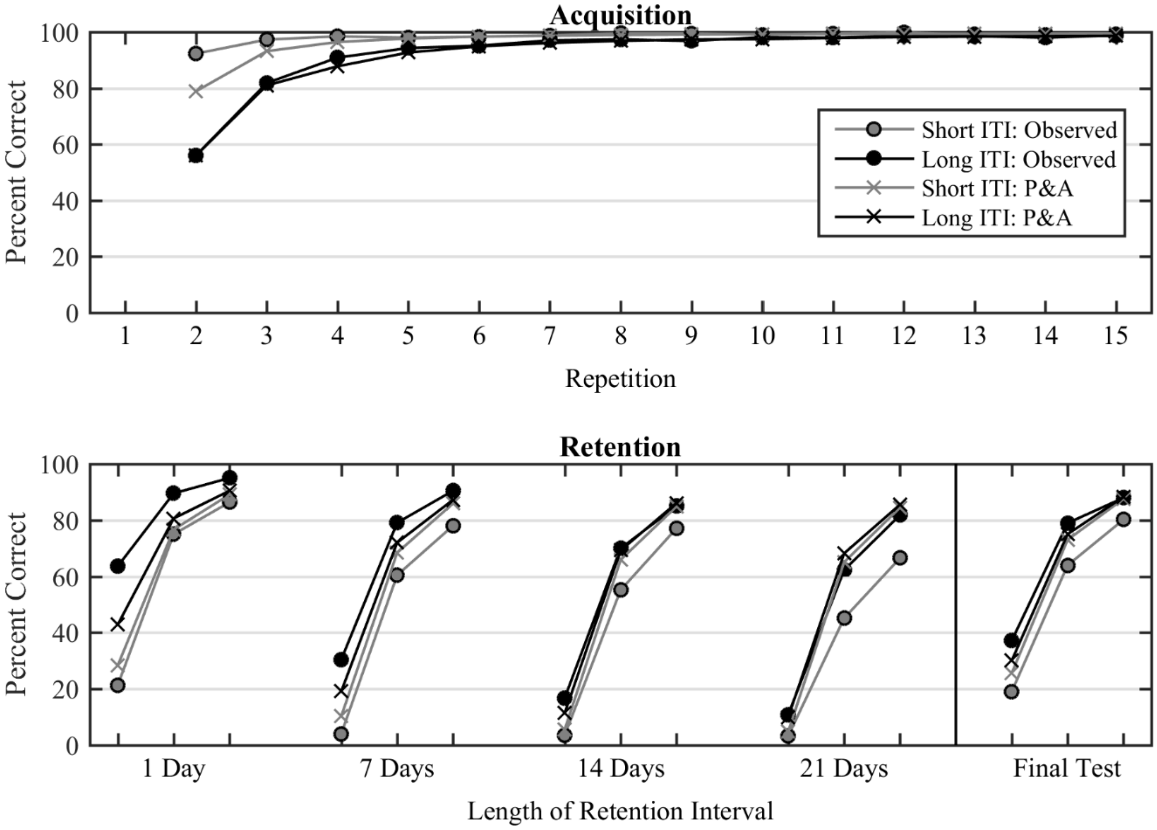

Figure 8. Observed data and best-fits of P&A for Experiment 1

P&A does not produce spacing-accelerated relearning because activation determines the probability of retrieval *and* the decay rates of instances subsequently added to memory. At the start of intermediate retention tests, activation of long ITI items equals or exceeds activation of short ITI items. Thus, long ITI items are slightly more likely to be retrieved, but their repetitions are stored with higher decay rates. Short ITI items effectively benefit more from re-exposure, eliminating or reversing the benefits of initial spacing across relearning trials.

Figure 9 shows the fit of PPE to data from the acquisition phase (top panel), and for the intermediate and final retention tests (bottom panel) of Experiment 1. The results were very similar for Experiment 2 (Supplementary Materials). The model accounted for all of the experiments' four key effects. First, PPE acquired items presented at a short ITI more quickly. In the model, elapsed time since earlier repetitions was less for items presented at a short ITI (*T* in Eq. 3), limiting the amount of forgetting that occurred across presentations of those items during the acquisition phase. Second, retention in PPE decreased with length of the retention interval because, more decay occurred with elapsed time between the acquisition phase and the intermediate retention test (*T* in Eq. 3). Third, retention in PPE was greater for items initially studied in a spaced manner. The durations of lags between repetitions (lag_j in Eq. 6) minimized decay rate in the model, which improved retention across the longer intervals between acquisition and test. Fourth and finally, PPE produced a relearning advantage for items initially studied in a spaced manner. The beneficial effects of spacing on decay rate in PPE (*d* in Eq. 3) were initially obscured after long retention intervals because the large elapsed time (*T*) caused performance to saturate near zero. The value of *T* quickly decreased across relearning trials, however, reflecting the contributions of repetitions from within the session. This allowed for the relative benefit of spaced practice, conferred through its impact on reducing decay rate, to reemerge.

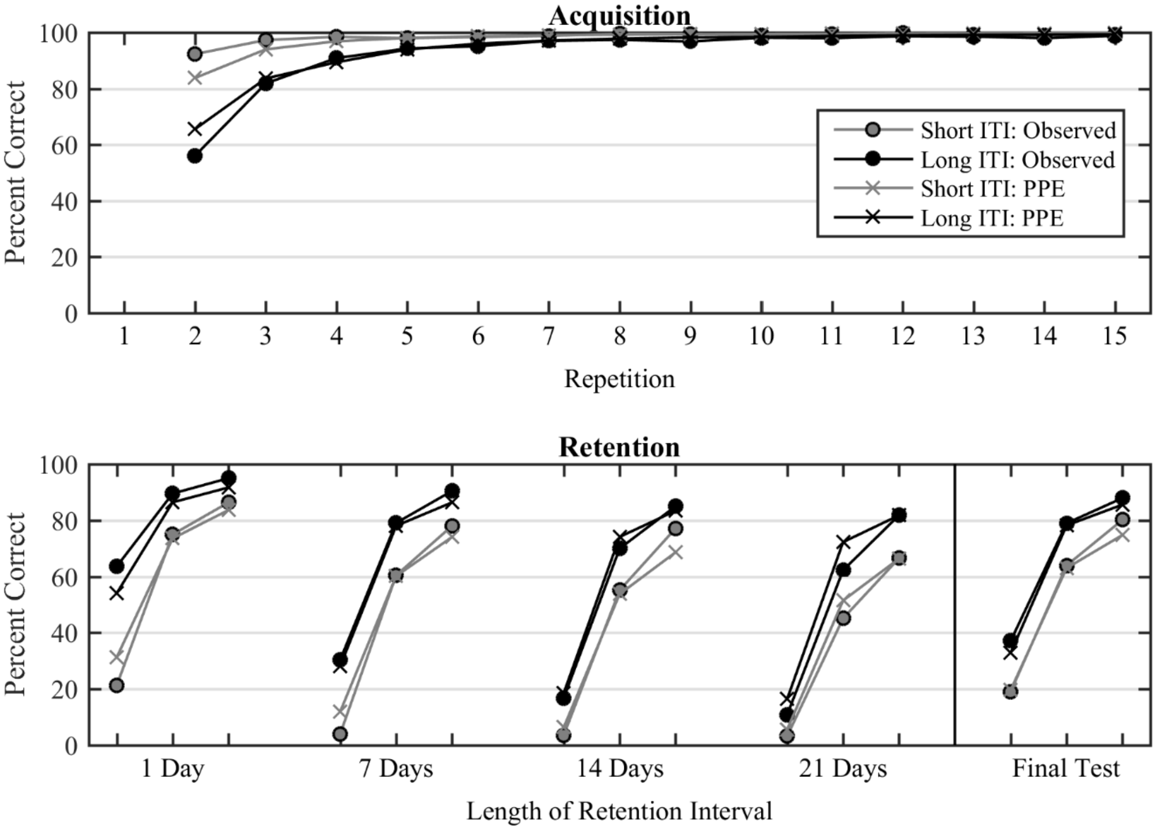

Figure 9. Observed data and best-fits of PPE for Experiment 1

Figure 10 shows the fit of SAM to the complete data from Experiment 1. The results from Experiment 2, which were comparable, are shown in the Supplementary Materials. The model accounted for three of the experiments' four key effects. First, SAM acquired items presented at a short ITI slightly faster because those items were more likely to remain in the STS at the time of their second and subsequent presentations. Additionally, even if the item had left the STS, SAM was more likely to retrieve it because there was little time for context to fluctuate during the brief intervals separating successive presentations of short ITI items. Notwithstanding the qualitative agreement, SAM produced a far smaller effect of ITI than was observed. Second, retention in SAM decreased with the length of the retention interval owing to the greater amount of contextual fluctuation that occurred from initial acquisition to test. Third, retention in SAM was greater for long versus short ITI items. When an item remained in the STS across successive repetitions, the item was not re-encoded in memory. Because short ITI items were more likely to remain in the STS, they effectively received less practice during initial acquisition. Additionally, even when a short ITI item left the STS and was successfully retrieved, the item was re-encoded with a largely overlapping set of contextual elements. Context fluctuates substantially across the long intervals separating initial acquisition from test. Following such intervals, the greater number of contextual elements encoded with long ITI items facilitated retention.

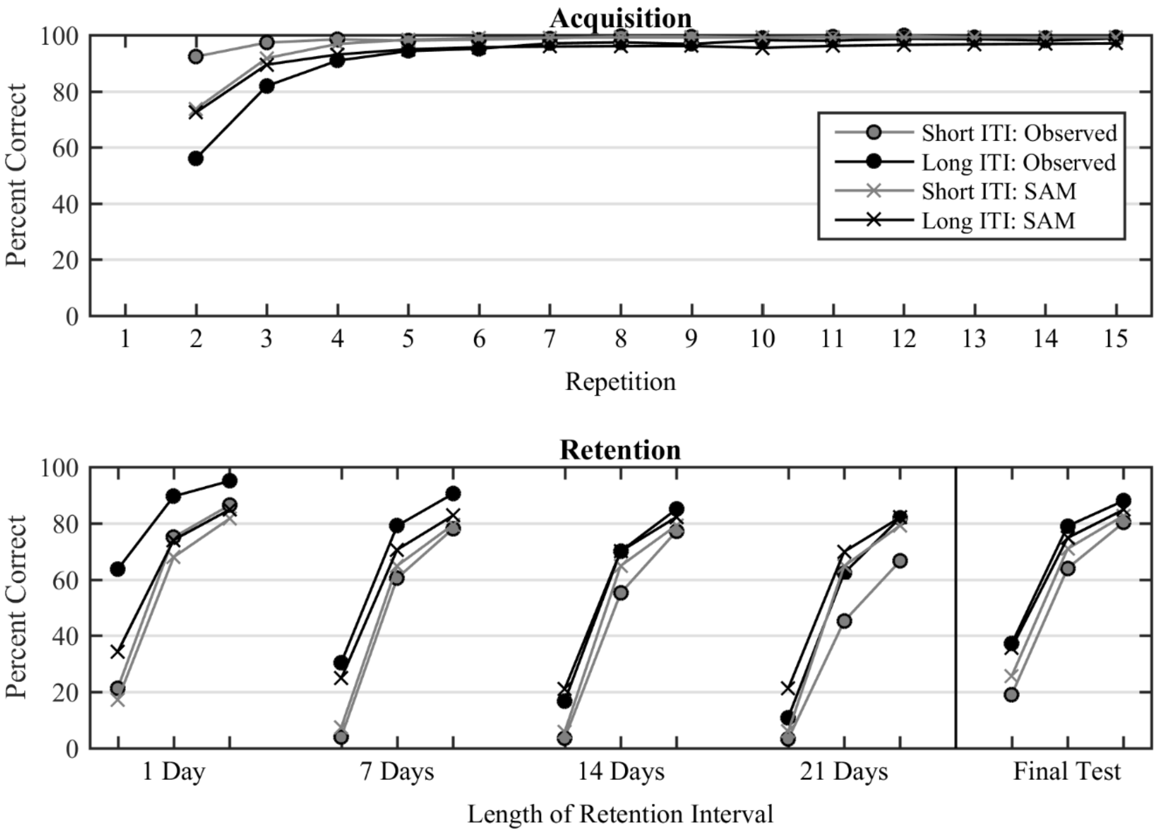

Figure 10. Observed data and best-fits of SAM for Experiment 1

SAM failed to produce a relearning advantage for items presented at long ITIs. This is a direct consequence of the retrieval-dependent update assumption—if an item is not successfully retrieved, a new trace is added to the LTS. The effects of initial study conditions, including the set of contextual elements encoded in the original trace, are lost when the new trace is added to the LTS. Thus, after a retrieval failure, all items will be relearned at the same rate.

Table 4 contains goodness of fit across the acquisition phase, intermediate retention tests, and final retention test for PPE, P&A, and SAM (treating percent correct as a whole number). Mean-square error during the acquisition phase was low for PPE, P&A, and moderate for SAM. Mean-square error was consistently lower for PPE than for P&A or SAM during all retention tests, however, boosting PPE's overall goodness of fit relative to P&A and SAM.

Table 4. Root mean-square error between observations and model fits to percent correct in Experiments 1 and 2

		Acquisition	**Intermediate Retention**	**Final Retention**	**Overall**
Experiment 1	P&A	2.8	9.0	6.4	6.4
	PPE	2.6	5.0	3.0	3.8
	SAM	5.1	9.9	4.6	7.4
Experiment 2	P&A	3.1	10.8	8.6	7.8
	PPE	2.3	7.1	2.0	4.9
	SAM	4.2	11.3	5.4	8.0

5.1.2. Model Discrimination

Do results from individual participants consistently favor one model? An answer to this question provides another test of model adequacy, but importantly, at the individual participant level. To address this question, we estimated the posterior probability of each model from the marginal likelihoods (Gelman et al, 2014; Wagenmakers, Morey, & Lee, 2016). These represent the probabilities of each of the three models in the evaluation set given the observations. From the posterior probabilities, we computed the log Bayes factor for the most probable model for each participant. Kass and Raftery (1995) proposed the following rule-of-thumb criteria for interpreting the log Bayes Factors: 1 to 3 denotes positive evidence for a model, 3 to 5 denotes strong evidence, and greater than 5 denotes very strong (i.e., definitive) evidence.

Figure 11 shows the posterior probabilities of the three models for all participants in Experiments 1 and 2. Most participants' data were more likely given PPE (Experiment 1: 17 of 22; Experiment 2: 20 of 25). Adopting Kass and Raftery's strictest criterion, there was very strong evidence for PPE for 32 participants, there was very strong evidence for P&A for 5 participants, and there was very strong evidence for SAM for 2 participants.

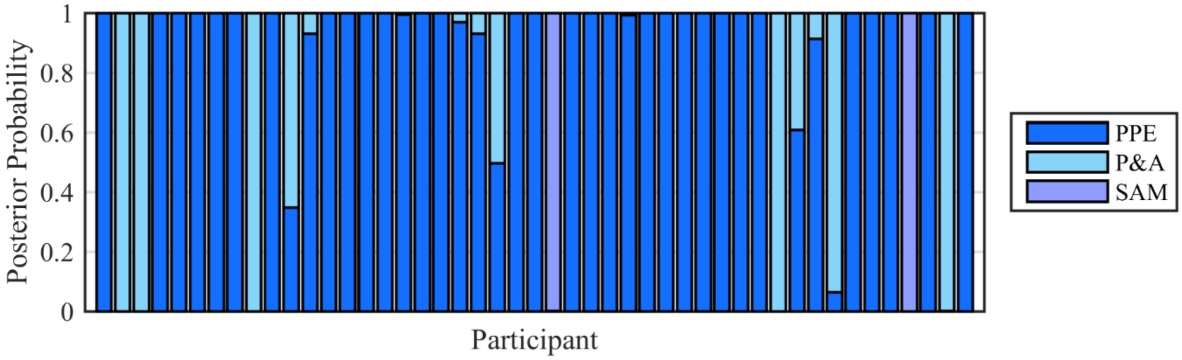

Figure 11. Posterior probabilities for all participants in Experiments 1 and 2.

The strength of evidence from the Bayes factors could indicate that PPE fit individuals exceptionally well, or conversely that P&A and SAM fit individuals poorly. Comparison of PPE's average fit with group level data (Figures 3 and 5) indicates the absence of systemic bias, which would appear as large differences between PPE and participants. However, averaging across model fits and participants could obscure large error residuals present for individuals. To address this issue, we calculated PPE's fit across all schedules and phases of the experiment for each individual. On average, PPE accounted for 89% of variance in individuals' performance in Experiment 1 (ranging from 78% to 93%) and 85% of variance in Experiment 2 (ranging from 45% to 95%). In sum, PPE provided a reasonable account of each participant's data, and of the three models, PPE provided the best account of most participants' data.

5.1.3. Model Prediction

The final question we asked was whether the models could accurately predict future performance. To answer this question, we estimated parameters based on subsets of each participant's data, and used the parameterized models to predict future performance. That is, we estimated parameters using data from acquisition through Day 1 retention, acquisition through Day 7 retention, and acquisition through Day 14 retention. In each case, we used the parameterized models to predict performance during the next session. Prequential methods to model evaluation (e.g., sequential prediction) are complementary to Bayesian methods (Shiffrin, Lee, Kim, & Wagenmakers, 2008). A model that captures only reliable structure in the data and not noise will generalize well to future data coming from the same source. Prequential methods also directly address a practical concern – to be used in an educational context, a model must accurately predict future, unseen performance data (Walsh et al., *submitted*). We limited our analysis to PPE. Given that P&A and SAM were unable to fit the relearning data, they would also fail to predict the data.

We fit PPE using all data through session *n*, and predicted performance in session *n+1* with the calibrated model. When calibrated using data from the acquisition phase and the 1 Day retention test, the model accurately predicted performance after the 7 day RI (Figure 12; Supplementary Table 4). When data from longer RIs were used for calibration, PPE's predictions for the next longest RI remained accurate (i.e., *accumulative* one-step look-ahead; Wagenmakers, Grunwald, & Steyvers, 2006). This indicates that given some retention data for parameter estimation, the model can accurately predict performance at least up to one week later. In fact, when calibrated using data from the 1 day RI, PPE predicted performance after 7, 14, and 21 days with minimal loss of accuracy (Supplementary Table 4).

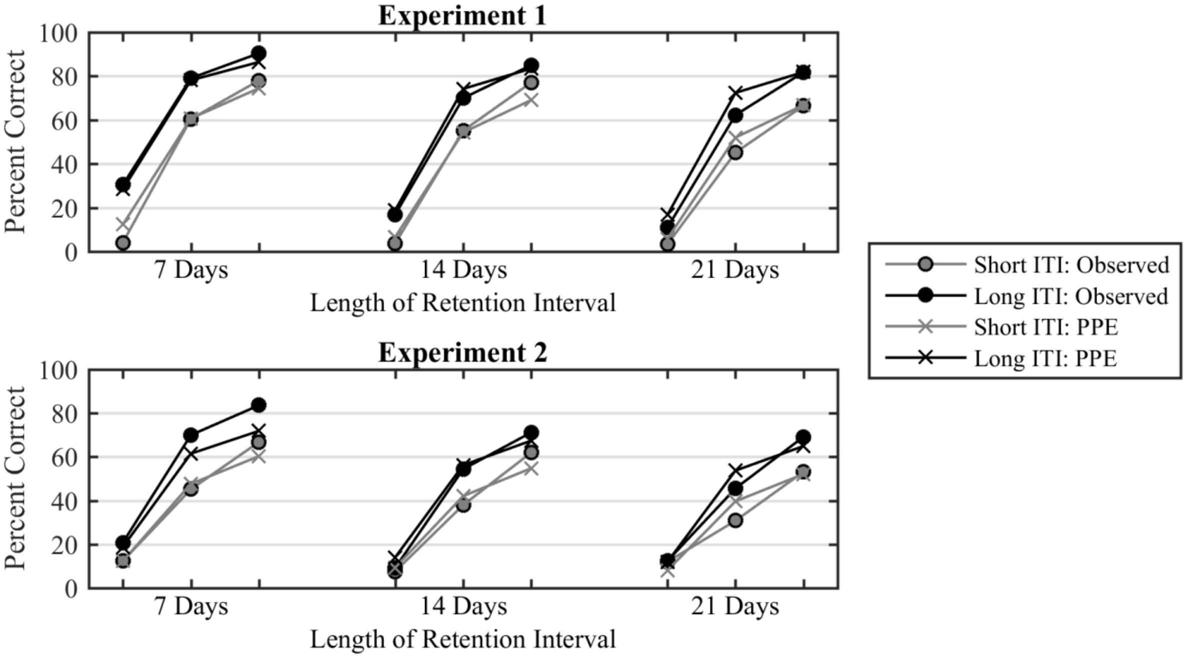

Figure 12. Observed data and predictions for PPE for Experiments 1 and 2.

The results contained in Figure 12 pertain to group-averaged performance. PPE's predictive accuracy holds at the level of individual participants, as well, (Supplementary Material). We calibrated PPE to individuals and generated out-of-sample predictions for the next retention test. Aggregating data across all participants, spacing conditions, retention intervals, and relearning trials, 49% observations were within 0.1 of predictions, 76% were within 0.2, and 89% were within 0.3.

5.2. Summary

The models captured several facets of the experimental results. All three produced an acquisition advantage for items presented at a short ITI and a retention advantage for items presented at a long ITI. PPE also produced a relearning advantage for spaced items, consistent with the experimental data, whereas P&A and SAM did not. In addition to giving a more complete account of the aggregate data, PPE was the more likely model for 79% (37/47) of participants, as found in our discrimination analysis. Lastly, in addition to adequately fitting data from acquisition, retention, and relearning, PPE made accurate out-of-sample temporal predictions at the level of the group and the individual. Collectively, these outcomes demonstrate the theoretical adequacy of PPE, and indicate that it could be used, for example, in an educational setting to predict retention and to decide when to review previously studied material.

6. General Discussion

The goal of this project was to evaluate the effects of spacing during initial acquisition on future relearning, and to evaluate three computational models of the spacing effect in light of those effects. The three models were motivated by different theoretical assumptions, they made qualitatively similar predictions regarding initial acquisition and retention, yet they made divergent predictions regarding relearning. In particular, PPE uniquely predicted that spacing would accelerate relearning. Two experiments confirmed the predictions of PPE, and subsequent model comparisons and analyses reinforced that conclusion.

Many experiments have varied the inter-trial intervals (ITIs) separating study repetitions. The consistent finding is that spacing slows acquisition but enhances retention. The absolute benefits of spacing gradually diminish, however, across retention intervals of one or more days. For example, Pashler et al. (2003) found that cued recall performance of Eskimo-English words after 1 day ranged from 20% to 50% depending on study spacing, Vaughn et al. (2016) found that cued recall of Swahili-English words after 1 week ranged from 10% to 30%, and Sobel et al. (2011) found that cued recall of unfamiliar English words after 5 weeks ranged from 10% to 20%. Spacing fortifies memory to a point, after which decay invariably drives performance back to zero.

An innovation of our experiments was to administer multiple cued recall trials with corrective feedback. This allowed us to measure the effects of spacing on retention *and* relearning following longer retention intervals (RIs). The effects were striking. After three weeks, memory performance for all items averaged below 10%, comparable with rates from other experiments that used similar materials and designs. After one or two relearning trials, however, items initially studied with a long ITI showed about a 20% absolute recall advantage relative to those studied with a short ITI. In fact, the relearning advantage for spaced items was present after *every* retention interval and on the final retention test once we controlled for retention differences. Of the items not successfully retrieved at the start of each test, items initially studied with a long ITI were relearned more quickly. These results go beyond mere model evaluation, and show that trial-level spacing enhances retention and exerts a powerful influence on the efficiency of relearning.

6.1. Theoretical Implications

What are the implications of these results with respect to the computational models of the spacing effect contained in Table 2 and their core mechanisms? PPE accounted for the effects of spacing on acquisition, retention, and relearning. In PPE, spacing causes memory traces to be stored with lower decay rates (i.e., *spacing-dependent trace decay*), enhancing retention. Critically, the effects of initial learning conditions (e.g., number of repetitions and spacing) in PPE persist across indefinitely long retention intervals, but may be obscured when elapsed time since previous repetitions is long (*T* in Eq. 5). Across relearning trials, elapsed time since the most recent trials decreases, allowing the effects of initial learning conditions to re-emerge.

PPE relates to the new theory of disuse (Bjork & Bjork, 1992). This theory distinguishes between storage strength and retrieval strength. Storage strength relates to how well an item is learned, and retrieval strength relates to how accessible it is from memory. For example, the zip code of a previous address may have high storage strength but low retrieval strength, whereas the name of a new acquaintance may have high retrieval strength but low storage strength. Storage strength increases monotonically with the number of opportunities to study or recall an item, and is never lost. Retrieval strength also increases when an item is studied or recalled, but is then gradually lost. The probability that an item can be recalled depends entirely on its retrieval strength. Storage strength, in turn, enhances the gain and slows the loss of retrieval strength after an item is studied or recalled.

The new theory of disuse predicts that relearning may restore the effects of initial learning conditions, including spacing. Although favorable initial learning conditions will increase an item's storage strength, a long enough retention interval will decrease the item's retrieval strength to the point that it can no longer be recalled. Subsequent opportunities to study or recall the item will boost its retrieval strength. These repetitions allow the item's storage strength, which does not decay, to potentiate gains and slow the future loss of retrieval strength. In sum, the new theory of disuse predicts that even in the absence of retrieval differences, items with high storage strength will be relearned more efficiently than those with low storage strength.

In PPE, the amount of practice along with the impact of the temporal distribution of practice on decay rate can be seen as contributing to an item's storage strength (*N* and *d* in Eq. 3). These quantities do not change with the mere passage of time. Elapsed time since practice contributes to an item's retrieval strength (*T* in Eq. 3). The greater the elapsed time, the less accessible the item becomes from memory. Subsequent opportunities to study or recall the item in PPE increase its retrieval strength by reducing elapsed time since the most recent repetitions, which are given greatest weight in the calculation of elapsed time (Eq. 5). This allows the impact of storage strength (*N* and *d* in Eq. 3) to re-emerge. In sum, a basic idea PPE and the new theory of disuse share is that differences in how strongly items are stored in memory can persist despite the absence of differences in how accessible they are from memory. *This subtle distinction between storage strength and retrieval strength, implicit in PPE, is the key feature of the model that allows it to account for spacing-accelerated relearning.*

In P&A, retrieval difficulty is inextricably linked to an item's activation (Eq. 1). Additionally, the decay rates associated with new instances of an item increase with the item's activation (i.e., *activation-dependent trace decay*). When practice is spaced, the activation of an item increases more gradually and new instances are stored with lower decay rates, enhancing retention (Eq. 2). This is consistent with the study phase retrieval hypothesis and the idea that more difficult retrievals produce greater learning gains. *The dual influence of activation on retrieval performance and trace decay is the key feature of P&A that causes it to fail to account for spacing-accelerated relearning.* If the effect of initial spacing, as measured by recall performance, is absent at the start of the test, relearning trials will produce similar gains for all items. If the effect of initial spacing remains, relearning trials will produce larger gains for the items that are most difficult to retrieve—that is, items that were initially massed. The paradoxical (and problematic) prediction of the model is that initial study conditions that enhance retention will hinder relearning. Other models in which a single source of memory strength drives recall performance and shapes subsequent learning gains would fail to account for spacing-accelerated relearning for the same reason.

In SAM, items are only re-encoded in memory once they have left the STS, and items are encoded along with contextual elements that fluctuate across time (i.e., *deficient processing* and *contextual variability*). When practice is spaced, item repetitions are more likely to be processed and they are encoded along with more varied contextual elements, enhancing retention. This is consistent with the deficient processing hypothesis and the contextual variability hypothesis. SAM also holds that contextual elements are only added to a trace if the trace is successfully retrieved (i.e., *retrieval dependent updates*). This is consistent with the study-phase retrieval hypothesis. *The retrieval-dependent update assumption is* *the key feature of SAM that causes it to fail to account for spacing-accelerated relearning*. Once a retrieval failure occurs, a new trace is added to the LTS. The set of contextual elements encoded in the original trace, which arise from the initial study conditions, are lost. This is inconsistent with our finding that initial spacing exerts a powerful influence on the efficiency of relearning items that are *not successfully retrieved* at the start of tests. Other models that contain the retrieval-dependent update assumption (e.g., MCM; Mozer et al., 2009) would also fail to account for spacing-accelerated relearning for this reason.

How could P&A or SAM be modified to capture spacing-accelerated relearning? The concept of activation-dependent trace decay is central to P&A, yet is fundamentally at odds with spacing-accelerated relearning. However, other aspects of the model could be modified to produce this effect. In ACT-R's complete activation equation (Anderson, 2007), the activation of an item is the sum of its base-level activation and spreading activation,

$$A_i = B_i + \sum_{j \in C} W_j S_{j,i}. \qquad (11)$$

A_i is the activation of item i, B_i is its base-level activation, C is the context defined as the set of retrieval cues, W_j is the attentional weight assigned to each cue, and $S_{j,i}$ is the strength of the association between each cue and item i. Assuming that context fluctuates over the course of the acquisition phase, items presented at a long ITI will be associated with more contextual elements (j in Eq. 11). This is an instantiation of the context variability hypothesis. Further assuming that relearning trials re-instantiate the initial study context, more activation will spread to spaced items. By this view, spaced items are not relearned more quickly. Rather, spaced items have more associations with the initial study context, which is re-instantiated during test. This is a substantial departure from the original P&A model. Simulations would be needed to determine whether the combination of activation-dependent decay and contextual variability can jointly produce spacing-accelerated relearning, and whether activation-dependent decay is still necessary once contextual variability is added to the model.

The retrieval-dependent update assumption is central to SAM, yet it prevents spacing-accelerated relearning. Removing the retrieval-dependent update assumption and updating memory traces after every retrieval attempt (c.f., Pavlik & Anderson, 2005) might allow SAM to produce spacing-accelerated relearning. However, this would have an undesirable side effect; the retrieval-dependent update assumption allows SAM to account for the non-monotonic effect of spacing interval on retention (Cepeda, Vul, Rohrer, Wixted & Pashler, 2008; Glenberg, 1976, Young, 1971). It is unclear whether and how SAM could be modified to simultaneously capture both effects.

Following Raajimakers (2003), we set the probability of retrieval to zero for any trace that was not retrieved at a certain trial. This simplifying assumption is justified based on three considerations: (1) The probability of retrieving a trace after a retrieval failure will be even lower than on the first test because the retention interval will be longer; (2) The probability of retrieving the trace will also be lower because an additional, new trace will be interfering with its retrieval; and (3) Empirically, in paradigms were one study trial is followed by two test trials, the probability of retrieving an item on the second test trial after responding incorrectly on the first is close to zero (Estes, 1960). Yet this is just a simplifying assumption, and allowing traces to remain retrievable following response errors might in some way influence relearning in the model.

What are the implications of these results with respect to theories of the spacing effect? The deficient processing hypothesis, the contextual variability hypothesis, and the study-phase retrieval hypothesis are general, and were not formed with consideration to the effects of spacing on relearning. The theories' predictions regarding efficiency of relearning, if any, are ambiguous. This is a limitation of descriptive theories of the spacing effect; because they are so abstract, they can be implemented in many different ways, and the manner in which they are implemented drives their predictions. This makes the theories difficult to falsify. Our experiments allow us to falsify specific implementations (i.e., assumptions) of these general theories. At some point, additional specificity is necessary to demonstrate the theory's continued relevance. The current data place additional constraints on theorizing.

Yet the results of our experiments have some implications for these theories, particularly the study phase retrieval hypothesis. If relearning is simply treated as a more sensitive measure of memory strength, then the finding that spacing enhances relearning after long intervals can be seen as an extension of the fact that spacing enhances retention after shorter intervals. In this sense, spacing-accelerated relearning is not inconsistent with study phase retrieval, contextual variability, or deficient processing.

When relearning is viewed as a process as well as a measure, however, the results challenge two aspects of the study-phase retrieval hypothesis. First, according to this hypothesis, the increase in memory strength produced by the presentation of an item is proportional to the difficulty of the retrieval (Benjamin & Tullis, 2010; Bjork, 1994; Hintzman, 2004). As shown by the P&A model, this leads to the prediction that spaced items, which may be easier to retrieve at the start of a retention test, will benefit less from relearning. This was not the case. Second, according to the study-phase retrieval hypothesis, a trace is only updated if it is successfully retrieved. As shown by SAM, this leads to the prediction that the effects of initial study will be lost once a retrieval failure occurs. Again, this was not the case. Although the study-phase retrieval hypothesis accounts for the effects of spacing on acquisition and retention, it is less clear how it could explain the effects of spacing on relearning.

The finding that initial study conditions impact relearning has some precedent. In a related experiment, de Jonge et al. (2014) examined the effects of repeated testing versus study on memory. Retrieval practice enhances memory relative to study; this is called the testing effect (for a review, see Roediger & Butler, 2011). de Jonge et al. (2014) replicated the testing effect. More importantly, they found that of the items forgotten after one week, those that had been initially tested rather than studied were relearned more quickly. Thus, evidence is accumulating that *spacing* and *testing*, two techniques known to improve retention, also both accelerate relearning.

Before turning from this work’s theoretical implications to its implications for training, two methodological details warrant attention. First, participants mastered massed items more quickly during the acquisition phase. Do the differential effects of overlearning for massed versus spaced items undermine our results? We think not. Participants reached similarly high levels of performance for all items by the end of the acquisition phase. Additionally, it is unclear why overlearning massed items would impair, rather than enhance, retention and relearning. Some experiments do use a learn-to-criterion procedure where items are dropped from study once they are successfully retrieved. This introduces another, potentially more problematic confound—the greater number of practice repetitions for spaced items. Based on these considerations, we chose to present the same number of repetitions in the short and long ITI conditions even though it led to overlearning of massed items.

Second, do the effects depend on details of the learning procedure that we employed; namely, one in which participants were tested and then restudied items during each trial. The test-restudy procedure is common in spacing effects experiments. Additionally, it is common in educational settings, giving it greater ecological validity. Aside from being widely used, the test-restudy procedure was necessary to evaluate the models’ predictions, for without a test we could not measure retention, and without restudy we could not measure relearning. Participants’ responses during the test portion of each trial provided a measure of retrieval success, which we used to evaluate the models. Additionally, the restudy (i.e., feedback) portion of each trial allowed us to investigate relearning. Without feedback, participants would have no way to relearn forgotten items. Notwithstanding the necessity of the procedure, the manner in which material is reviewed (study, test, or a combination of both) does affect retention. Retrieval practice is superior to study (Roediger & Butler, 2011). Some experiments have attempted to separate the effects of trial spacing from review method (study versus test) using factorial designs. Of those experiments, two reported additive effects of spacing and review method (Carpenter, Pashler & Cepeda, 2009; Pyc & Rawson, 2012), and one reported an enhanced spacing effect in the testing condition (Rawson, Vaughn, & Carpenter, 2015). Future experiments are needed to understand precisely how spacing and testing jointly affect relearning, though this prior work indicates that eliminating testing would likely result in quantitative, rather than qualitative changes to the pattern of results.

6.2. Educational Practice

The educational implications of this work are multifaceted. Most experiments of the spacing effect involve a single session with retention intervals ranging from seconds to minutes (for a review and exceptions, see Cepeda et al., 2006). Likewise, most theories of the spacing effect were proposed to account for learning and retention within one experimental session (Bahrick et al., 1993). To support educational practice, experiments, theories, and models of the spacing effect must scale up to educationally-relevant timescales ranging from days to years (c.f. Anderson, 2002; Bahrick, 1979; Cepeda et al., 2008; Walsh et al., *submitted*). In the current experiment, we studied learning, retention, and relearning across timescales ranging from one day to three weeks. Results from the longest retention interval provided mixed support for the utility of spacing as a study strategy. On the one hand, retention was initially equally low for massed and spaced items. On the other hand, the relative benefits of spacing reemerged in terms of efficiency of relearning.

Much of what students learn is forgotten (Bahrick, 1979), yet memory is enhanced when studied material is re-experienced. The repeated cycle of learning, loss, and relearning may establish durable long-term memories (Bahrick, 1979; Rawson, Vaughn, Walsh, & Dunlosky, *submitted*). With this in mind, relearning is an important criterion for judging the educational value of an intervention. Durable learning is important, but so too is efficient relearning (Rawson & Dunlosky, 2011). Nelson (1985) noted this, saying that teaching may be beneficial by allowing students to relearn information relatively quickly. The results of our experiments establish that spaced retrieval practice is one way to maximize these benefits.

Similarly, the medical community recognizes that psychomotor skill decay is inevitable. This has led to the identification of optimal intervals to administer refresher training of basic life support skills such as cardiopulmonary resuscitation (CPR; Oermann, Kardong-Edgren, & Odom-Maryon, 2011). In a variant of this called just-in-time (JIT) training, proficiency is restored by providing refresher training immediately before a skill is applied (Barnes, 1998). Educators also recognize the unavoidable effects of decay on memory performance, leading to an active body of research on when to optimally schedule review sessions (Carpenter et al., 2012). These examples demonstrate the utility of combining initial training with relearning. In medicine and education alike, information must be periodically reviewed to restore and maintain proficiency. The current work reinforces the finding that information, once studied, is relearned more quickly. It further shows that the manner in which information was initially studied affects how quickly it is relearned. Thus, this work is informative with respect to the goal of delivering initial study in a manner that facilitates later relearning.

Laboratory studies have expanded beyond single session designs where spacing is manipulated by varying the number of trials between item repetitions, to multi-session designs where elapsed time between sessions on separate days is manipulated. These studies also consistently reveal a spacing advantage; as elapsed time between sessions increases, so too does long-term retention. The benefits of studying information on multiple occasions are impressive. For example, Cepeda et al. (2008) found that when participants studied trivia facts during two sessions, retention one week later ranged from 75% to 100% depending on the number of days between the study sessions, and performance two months later ranged from about 30% to 60%. Still, given enough time, retention becomes low (e.g., 20% after one year in Cepeda et al., 2008). An interesting direction for future research is to vary the spacing between two sessions on separate days, and to administer multiple relearning trials during the final retention test. The benefits of session spacing (like the benefits of trial spacing in the current experiments), though small across long enough retention intervals, may reemerge during relearning.

Our conclusion that spacing facilitates relearning may seem at odds with some existing multi-session experiments (Rawson & Dunlosky, 2013; Rawson et al., *submitted*). In those experiments, trial-level spacing was manipulated during an initial study session. All items were then studied with equal trial-level spacing during two or more relearning sessions. With each successive relearning session, the effects of the initial spacing manipulation decreased. This has been called the *relearning override effect* (Rawson et al., *submitted*; Vaughn, Dunlosky, & Rawson, 2016). Our conclusion is not in tension with this result. Spacing can facilitate relearning. At the same time, relearning, which itself is a powerful study technique (Bahrick, 1979; Rawson & Dunlosky, 2011), may overwhelm the effects of initial spacing when it is repeatedly administered. The multi-faceted nature of learning is one of the reasons its optimization remains elusive.

Another educational implication of our research relates to the robustness of spacing-accelerated learning in different settings and with different populations. We replicated the current experiment with two groups of participants and in two different settings: first with a heterogeneous group of college-aged participants in a well-controlled laboratory setting, and then with a more diverse group of participants from throughout the United States and with little control over the setting. Consistent with previous studies, performance of participants in the MTurk version of the experiment was lower on average (Goodman, Cryder, & Cheema, 2013; Simcox & Fiez, 2014). Most importantly, spacing affected learning, retention, and relearning identically in both versions of the experiment.

In evaluating 10 different learning techniques, Dunlosky et al. (2013) heavily weighted evidence of their effectiveness in educational contexts. Relatedly, Dempster (1988) cited demonstration of phenomena in a school-like setting as essential to their eventual application. A robust laboratory finding may fail to replicate in educational settings for various reasons: for example, if it is dependent on specific learning conditions or student characteristics (Dunlosky et al., 2013). The benefits of spaced practice on retention have been demonstrated in laboratory studies, and in real and simulated classrooms (for reviews, see Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Dunlosky et al., 2013). Although we did not specifically examine the benefits of spaced practice on relearning in a classroom, the consistency of the results across two different groups of participants in different settings indicate that they are not heavily dependent upon learning conditions or student characteristics. A potential direction for future research is to examine the impact of spacing on relearning in an actual educational setting.

The final implication of this research concerns the use of computational models to account for the effects of spaced practice and other educational variables on learning and retention. Various models have been proposed to predict the effects of certain variables on learning *or* retention. In actuality, these are ongoing, dynamic, and competing processes. To fully appreciate the implications of an instructional intervention, it is necessary to consider its impact on (1) learning, (2) forgetting, and (3) relearning. This is consistent with the call to study the effects of instructional interventions on initial performance along with long-term retention to identify potential tradeoffs (Schmidt & Bjork, 1992; Soderstrom & Bjork, 2015). Computational models, like theories, should account for the effects of educational variables on the complete set of processes.

By administering massed and spaced practice during acquisition, and measuring performance after multiple retention intervals, it was possible to examine the impact of spacing on learning, retention, and relearning. Further, it was possible to test whether PPE could simultaneously account for the impact of spacing on the complete set of processes. PPE fit all trends in the group-level data. The success of PPE indicates that it can be used to quantify the costs (i.e., time in training) and risk (i.e., proficiency over time) associated with different practice schedules across the complete training lifecycle. A validated computational model like PPE allows one to answer nuanced questions such as how a fixed number of practice repetitions should be spread to maximize retention, how much practice is needed to achieve a target level of proficiency, and how many relearning trials are needed to restore proficiency. It is often impractical to answer these questions by gathering student data for all levels and combinations of educational variables (Koedinger, Booth, & Klahr, 2013).

The accuracy of PPE's out-of-sample temporal predictions at the group level indicate that it can be used to identify fixed practice schedules that achieve the desired learning outcomes for the majority of students. Characteristics of the individual may interact with training variables however, making fixed schedules more or less advantageous for different students. PPE's predictive power at the level of individuals indicates that it can also be used to anticipate when a specific student will no longer be able to remember a certain item. This information can be used to deliver personalized review in the place of fixed practice (Khajah, Lindsey, & Mozer, 2014; Lindsey et al., 2014; Pavlik & Anderson, 2008; Walsh et al., *submitted*). Given recent evidence favoring adaptive over fixed practice schedules (Mettler, Massey, &

Kellman, 2016), applying PPE at the level of individuals would likely achieve the most positive learning outcomes.

6.3. Conclusion

What factors foster relearning? Conditions that optimize acquisition might be expected to optimize retention, but this is not always true (Schmidt & Bjork, 1992; Soderstrom & Bjork, 2015). Likewise, conditions that optimize retention might be expected to optimize relearning. Again, this is an empirical question. In our experiments, we found that spacing improved retention *and* relearning. The finding of spacing-accelerated relearning was most apparent after long retention intervals despite the initial absence of differences in memory for spaced versus massed items. To echo and extend Schmit and Bjork's message that initial learning and retention must be considered together (1992), the goal of training, and the focus of educational research, should be expanded to include relearning.

Acknowledgements

The views expressed in this paper do not reflect the official views or opinions of the United States or the United States Air Force. Funding for this research was provided by the U.S. Air Force Research Laboratory early career award to T.J., and by grant 13RH15COR from the Air Force Office of Scientific Research (AFOSR). The ideas and data in this manuscript were not disseminated elsewhere.

References

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science, 26*, 85-112.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe.* New York, NY: Oxford University Press.

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

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in Instructional Psychology: Educational Design and Cognitive Science*, Vol. 5. Mahwah, NJ: Erlbaum.

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

Barnes, B. E. (1998). Creating the practice-learning environment: using information technology to support a new model of continuing medical education. *Academic Medicine*, *73*, 278-81.

Bates D. M., Maechler M., & Bolker B. (2014). lme4: Linear mixed–effects models using S4 classes. R package version 0.999999–0. Retrieved from http://CRAN.R-project.org/package=lme4

Begg, I., & Green, C. (1988). Repetition and trace interaction: Superadditivity. *Memory and Cognition, 16*, 232–242.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*, 228-247.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31*, 635-650.

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: evidence from once-presented words. *Memory, 6*, 37-65.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*, 760–771

Carpenter, S., Cepeda, N., Rohrer, D., Kang, S. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369-378.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

de Jonge, M., Tabbers, H. K., & Rikers, R. M. (2014). Retention beyond the threshold: Test-enhanced relearning of forgotten information. *Journal of Cognitive Psychology*, *26*, 58-64.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63-147.

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

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4-58.

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.

Estes, W. K. (1960). Learning theory and the new "mental chemistry". *Psychological Review*, *67*, 207–223.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014) *Bayesian Data Analysis (Third Edition)*. Boca Raton, FL: CRC 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.

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

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213-224.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 371-377.

Greeno, J. G. (1964). Paired-associate learning with massed and distributed repetitions of items. *Journal of Experimental Psychology*, *67*, 286-295.

Groninger, L. K., & Groninger, L. D. (1980). A comparison of recognition and savings as retrieval measures: A reexamination. *Bulletin of the Psychonomic Society*, *15*, 263-266.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77–97). Potomac, MD: Erlbaum.

Hintzman, D. L. (2004). Judgment of frequency vs. recognition confidence: Repetition and recursive reminding. *Memory and Cognition, 32*, 336–350.

Hintzman, D. L., & Rogers, M. K. (1973). Spacing effects in picture memory. *Memory and Cognition, 1*, 430–434.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of*

Consumer Research, 30, 138–149.

Jastrzembski, T. S., & Gluck, K. A. (2009). A formal comparison of model variants for performance prediction. *Proceedings of the International Conference on Cognitive Modeling*, Manchester, UK.

Jost, A. (1897). Die Assoziationsfestigkeit in ihrer Abha¨ngigkeit von der Verteilung der Wiederholungen [The strength of associations in their dependence on the distribution of repetitions]. *Zeitschrift fu¨r Psychologie und Physiologie der Sinnesorgane, 14*, 436–472.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. *Topics in Cognitive Science*, *6*, 157-169.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*, 935-937.

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

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*, 639-647.

MacLeod, C. M. (1976). Bilingual episodic memory: Acquisition and forgetting. *Journal of Verbal Learning & Verbal Behavior, 15,* 347–364.

MacLeod, C. M. (1988). Forgotten but not gone: Savings for pictures and words in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 195-212.

Maner, J. K. (2014). Let's put our money where our mouth is. If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives in Psychological Science, 9*, 343-351.

Mazza, S., Gerbier, E., Gustin, M. P., Kasikci, Z., Koenig, O., Toppino, T. C., & Magnin, M. (2016). Relearn Faster and Retain Longer Along With Practice, Sleep Makes Perfect. *Psychological Science*, *27*, 1321-1330.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*, 11-38.

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.

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A Comparison of Adaptive and Fixed Schedules of Practice. *Journal of Experimental Psychology: General.*

Moulton, C. A. E., Dubrowski, A., Macrae, H., Graham, B., Grober, E., & Reznick, R. (2006). Teaching surgical skills: what kind of practice makes perfect? A randomized, controlled trial. *Annals of Surgery, 244*, 400–409.

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1321-1329). La Jolla, CA: NIPS Foundation.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47,* 90-100.

Nelson, T. O. (1978). Detecting small amounts of information in memory: Savings for nonrecognized items. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 453-468.

Nelson, T. O. (1985). Ebbinghaus's contribution to the measurement of retention: Savings during relearning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 472–479.

Oermann, M. H., Kardong-Edgren, S. E., & Odom-Maryon, T. (2011). Effects of monthly practice on nursing students' CPR psychomotor skill performance. *Resuscitation*, *82*(4), 447-453.

Paolacci, G., & Chandler, J. (2014). Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184-188.

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

Pavlik, 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.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*, 101-117.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437–447.

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 737-746.

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

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*, 283-302.

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*, 1113-1129.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & cognition*, *43*(4), 619-633.

Rawson, K., Vaughn, K., Walsh, M., & Dunlosky, J. (*submitted*). Successive relearning: The next frontier for educationally relevant memory research.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20-27.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209–1224.

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.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology, 19*, 107-122.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248-1284.

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*, 95-111.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76-80.

Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, *17*, 77-88.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*, 763-767.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance an integrative review. *Perspectives on Psychological Science*, *10*, 176-199.

Thorndike, E. L. (1912). The curve of work. *Psychological Review, 19*, 165–194.

Toppino, T. C., & Gerbier, E. (2014). About practice: repetition, spacing, and abstraction. *The Psychology of Learning and Motivation*, *60*, 113-189.

Vaughn, K. E., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory & Cognition*, 1-13.

Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 796-800.

Wagenmakers, E.-J., Grunwald, P. D., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50*, 149-166.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (*submitted*). Evaluating the theoretical adequacy and applied potential of computational models of the spacing effect.

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.

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