Evaluating the Theoretic Adequacy and Applied Potential of Computational Models of the Spacing Effect

Matthew M. Walsh,\textsuperscript{a} Kevin A. Gluck,\textsuperscript{b} Glenn Gunzelmann,\textsuperscript{b} Tiffany Jastrzembski,\textsuperscript{b} Michael Krusmark\textsuperscript{b,\textcopyright}

\textsuperscript{a}The RAND Corporation
\textsuperscript{b}Air Force Research Laboratory, Wright-Patterson Air Force Base
\textsuperscript{c}L-3 Communications, Wright-Patterson Air Force Base

Received 16 August 2016; received in revised form 26 September 2017; accepted 26 January 2018

Abstract

The spacing effect is among the most widely replicated empirical phenomena in the learning sciences, and its relevance to education and training is readily apparent. Yet successful applications of spacing effect research to education and training is rare. Computational modeling can provide the crucial link between a century of accumulated experimental data on the spacing effect and the emerging interest in using that research to enable adaptive instruction. In this paper, we review relevant literature and identify 10 criteria for rigorously evaluating computational models of the spacing effect. Five relate to evaluating the theoretic adequacy of a model, and five relate to evaluating its application potential. We use these criteria to evaluate a novel computational model of the spacing effect called the Predictive Performance Equation (PPE). Predictive Performance Equation combines elements of earlier models of learning and memory including the General Performance Equation, Adaptive Control of Thought—Rational, and the New Theory of Disuse, giving rise to a novel computational account of the spacing effect that performs favorably across the complete sets of theoretic and applied criteria. We implemented two other previously published computational models of the spacing effect and compare them to PPE using the theoretic and applied criteria as guides.

Keywords: Spacing effect; Computational model; Education; Predictive Performance Equation

1. Introduction

The acquisition and retention of knowledge are impacted by multiple factors, including amount of practice, elapsed time since practice, and distribution of practice over time...
The third factor, temporal distribution of practice, is central to memory research on the spacing effect. This research has shown that separating practice repetitions by a delay—that is, *spacing*—enhances retention (Ebbinghaus, 1885/1964; Jost, 1897; Thorndike, 1912). The effect seems counterintuitive because the opposite approach of condensing practice into a short period of time—that is, *massing*—facilitates acquisition. Yet the spacing effect is one of the most widely replicated results in psychology research (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), and its potential implications for education and training are substantial (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012).

Many theories have been proposed to account for the spacing effect. One set appeals to the notion of *deficient processing* when practice is not sufficiently spaced (Greene, 1989; Hintzman, 1974). When people experience a repetition after a short delay, they dedicate less attention or effort to encode the second presentation. Inserting space between repetitions ensures that people can adequately attend to and process each repetition. A second set of theories is based on the idea of *encoding variability* (Estes, 1955; Glenberg, 1979; Melton, 1970). According to these, each practice repetition results in the encoding of the stimulus along with contextual elements. Context changes over time. As the space between repetitions increases, the redundancy of encoded contextual elements decreases. This improves memory performance at test because the storage of more varied contextual information provides more routes through which the target stimuli can be accessed. A final set of theories propose that the spacing effect arises from a *study-phase retrieval* process (Benjamin & Tullis, 2010; Bjork, 1994; Hintzman, 2004). By this view, retrieval of past presentations of an item strengthens the original memory trace. The difficulty of retrieval influences the change in strength: The more difficult the retrieval, the greater the boost. Memory updates are contingent, however, upon the successful retrieval of the item (i.e., the retrieval-dependent update assumption; Benjamin & Tullis, 2010; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009).

These theories have motivated considerable debate. Evaluating them is both enabled and complicated by the diversity of the memory phenomena associated with the spacing effect (for reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Küpper-Tetzel, 2014; Pavlik & Anderson, 2005; Raaijmakers, 2003; Toppino & Gerbier, 2014). The complexity of those findings not only impedes progress in our scientific understanding of the spacing effect, but also impedes the application of spacing effect research in education. Dempster (1988) cited “incomplete understanding of the psychological basis of the spacing effect” as a key barrier to apply spacing research to education. Anderson and Schunn (2000) make this point more generally, saying “To be able to rigorously understand what the implications are of cognitive psychology research one needs a rigorous theory that bridges the gap between the detail of the laboratory experiment and the scale of the educational enterprise” (p. 1).

Computational modeling is one approach to address both the theoretic and practical concerns associated with research on the spacing effect. In terms of theory, computational models offer a way to unify the ever-expanding collection of empirical results related to the spacing effect. In terms of application, computational models provide the kind of rigor that...
Anderson and Schunn (2000) call for and can lead to technologies that improve education (Anderson, Corbett, Koedinger, & Pelletier, 1995; Mettler, Massey, & Kellman, 2016; Pavlik & Anderson, 2008; for a review, see Walsh & Lovett, 2016). The insights gained from theory and application are mutually beneficial. Understanding the mechanisms that impact performance in laboratory studies is important because they affect behavior in the world as well. Conversely, if a model’s predictions are validated in an educational or training setting, those data provide further support for the computational theory.

In this paper, we propose a general set of criteria for evaluating the theoretic adequacy and the applied potential of computational models of the spacing effect. We focus on the spacing effect because it is well documented, extremely relevant to education and training, and the focal point of a family of computational models (e.g., Mozer et al., 2009; Pavlik & Anderson, 2005; Raaijmakers, 2003). Establishing a set of theoretic criteria will facilitate comparison of the computational models, their core mechanisms, and the accounts of the spacing effect that they relate to. Establishing a set of applied criteria will narrow the gap between computational models and educational and training practices.

After proposing theoretic and applied criteria for evaluating models of the spacing effect, we present a new implementation of a computational model called the Predictive Performance Equation (PPE; Jastrzembski & Gluck, 2009). Predictive Performance Equation is an extension of a mathematical model of knowledge acquisition and retention called the General Performance Equation (Anderson & Schunn, 2000), and PPE embodies elements of the New Theory of Disuse (Bjork & Bjork, 1992), a general theory of learning and memory not specifically tied to the spacing effect. Predictive Performance Equation was motivated, in part, by limitations of existing computational models in accounting for the full range of spacing-related phenomena. Although most models explain one or a small number of spacing-related phenomena, none have been evaluated against as broad a set of results as the criteria and comparative evaluation we provide here. Additionally, the implications of computational models of the spacing effect for education and training are rarely considered (but see Khajah, Lindsey, & Mozer, 2014; Lindsey, Shroyer, Pashler, & Mozer, 2014; Pavlik & Anderson, 2008). To that end, we evaluate PPE according to the complete set of applied criteria as well. To facilitate comparison, we evaluate two other models of the spacing effect alongside PPE. The first is an extension of ACT-R (Pavlik & Anderson, 2005), and the second is based on the Search of Associative Memory (SAM) model (Raaijmakers, 2003). These are among the most widely studied, cited, and used models of the spacing effect, they contain enough detail to make testable predictions, and they contain different core mechanisms that relate to different accounts of the spacing effect. In essence, they form a meaningful set of competitors for gauging PPE’s success.

2. Model evaluation criteria

The memory literature is replete with experiments that reveal nuances of the spacing effect (for recent reviews, see Cepeda et al., 2006; Delaney et al., 2010). These comprise
a set of criteria for evaluating the theoretic adequacy of computational cognitive models. In parallel, the education literature underscores key issues that must be addressed to enable practice recommendations (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Koedinger, Booth, & Klahr, 2013; Mozer & Lindsey, 2017). These form the basis for evaluating the practical utility of computational cognitive models. We elaborate on both sets of criteria in turn.

2.1. Theoretic evaluation

To identify theoretic criteria for evaluating computational models of the spacing effect, we first consulted a set of review papers (Benjamin & Tullis, 2010; Cepeda et al., 2006; Delaney et al., 2010; Küpper-Tetzel, 2014; Pavlik & Anderson, 2005; Toppino & Gerbier, 2014). We then expanded our search to include a broad range of journal articles published on the effects of spacing. Across review papers and the psychological literature more generally, five phenomena consistently emerged (Table 1).

2.1.1. Role of spacing on retention

The acquisition and retention of knowledge is affected by the temporal distribution of practice (Ebbinghaus, 1885/1964; Jost, 1897; Thorndike, 1912). Many experiments have demonstrated this by manipulating elapsed time between item repetitions (for a review, see Cepeda et al., 2006). When study time devoted to a single item is not interrupted, learning is massed. Alternatively, when measurable time or intervening items separate study opportunities, learning is spaced. Massed practice typically has the desirable effect of accelerating learning and the undesirable effect of accelerating forgetting, whereas spaced practice has the opposite effects, slowing initial learning and enhancing retention. Historically, the term “spacing effect” describes this phenomenon.

The spacing effect is extremely general. It has been demonstrated in memory tasks involving pictures or words (Hintzman & Rogers, 1973; Janiszewski, Noel, & Sawyer, 2003), in the acquisition of perceptual-motor skills (Lee & Genovese, 1988; Moulton et al., 2006; Stafford & Dewar, 2014), and in the attainment of educationally relevant abilities (Rohrer & Taylor, 2006; Seabrook, Brown, & Solity, 2005). The spacing effect holds across timescales ranging from seconds to years (Bahrick, 1979; Cepeda et al., 2006). The effect has been observed in a multitude of laboratory experiments and more recently in ecologically valid educational settings (Carpenter et al., 2012). Finally,

Table 1
Experimental phenomena for evaluating computational models of the spacing effect

<table>
<thead>
<tr>
<th></th>
<th>Role of spacing on retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Relationship between retention interval and optimal spacing interval</td>
</tr>
<tr>
<td>2</td>
<td>Increased benefit of spacing with amount of practice</td>
</tr>
<tr>
<td>3</td>
<td>Attenuation of spacing effect with re-learning</td>
</tr>
<tr>
<td>4</td>
<td>Superadditive learning gains from repetition</td>
</tr>
</tbody>
</table>
children, young adults, and older adults all show the spacing effect, as do different animal species (Delaney et al., 2010).

2.1.2. Relationship between retention interval and optimal spacing interval

The spacing effect refers to enhanced retention for spaced relative to massed practice repetitions. Researchers have also manipulated the duration of time between study opportunities—that is, the interstudy interval (ISI)—to produce different levels of spacing (Fig. 1). In these experiments, the amount of time between study opportunities varies by condition, and performance is measured after a fixed retention interval (RI) following the final study opportunity. When two schedules with different amounts of space between repetitions are compared (i.e., spacing is varied in a graded manner), the results of degree of spacing on retention are called lag effects rather than spacing effects.

Some degree of spacing usually improves performance. However, the duration of the ISI has a non-monotonic effect on retention. Increasing the time between study opportunities is beneficial to a point, after which it begins to impair final retention (Benjamin & Tullis, 2010). This is seen in cued- and free-recall paradigms (Peterson, Wampler, Kirkpatrick, & Saltzman, 1963; Young, 1971), and in experiments that vary the inter-trial interval (i.e., minutes) or the inter-session interval (i.e., days) between item repetitions (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Toppino, Hara, & Hackman, 2002). The retention function has an inverted U-shape and is right-skewed (Cepeda et al., 2008; Peterson et al., 1963; Young, 1971). As the ISI increases, final retention rises sharply before peaking, after which it gradually drops. The steepness of the curve varies with the length of the RI, such that the effects of ISI, in absolute terms, diminish as performance approaches floor after very long RIs (Cepeda et al., 2008).

![Fig. 1. Spacing schedules created by varying the interstudy interval (ISI) between study opportunities and retention interval before the final retention test. Multi-session schedules are created by introducing additional study sessions with fixed or varying ISIs before final test.](image-url)
Many studies also vary the duration of the RI (Fig. 1). These studies show that the length of the ISI interacts with the length of the RI in determining memory performance: Massed practice enhances retention for very short RIs, and spaced practice enhances retention for long RIs. Stated differently, as the RI increases, so too does the value of the ISI that maximizes retention (Cepeda et al., 2008; Glenberg, 1976). Cepeda et al. (2008) refer to the relationship between the RI, the ISI, and retention as a “temporal ridgeline of optimal retention.” The optimal ISI increases with the RI, but as a decreasing ratio of it. This interaction is obtained in studies that occur over a brief period of time (i.e., a single experiment session; Atkinson & Shiffrin, 1968; Balota, Duchek, & Paullin, 1989; Glenberg, 1976), and over longer durations (i.e., two or more sessions separated by days and months; Cepeda et al., 2008; Glenberg & Lehman, 1980; Küpper-Tetzel & Erdfelder, 2012).

2.1.3. Increased benefit of spacing with amount of practice

The benefits of spacing accumulate with amount of practice. In a study that demonstrated this interaction, Pavlik and Anderson (2005) varied the number of study repetitions and the spacing between them in a paired-associate learning task. As the number of repetitions increased, the effect of spacing on final retention measured after 1 and 7 days became larger. The same interaction has been observed in free-recall paradigms (Shaughnessy, Zimmerman, & Underwood, 1972; Underwood, 1970). Likewise, raising the initial criterion level, which increases the amount of practice, also enhances the spacing effect (Pyc & Rawson, 2009; Rawson & Dunlosky, 2013).

This phenomenon bears on the issue of overlearning—that is, providing continued retrieval practice after a criterion of one (or more) errorless trials has been reached. Studies of verbal learning and skill acquisition consistently support the effectiveness of overlearning in enhancing retention (Driskell, Willis, & Copper, 1992). The potency of spacing is evidently enhanced by overlearning. The appearance of the spacing-by-practice interaction, as with all other spacing-related effects, depends on final retention not reaching ceiling or floor. Yet the general point is that additional spaced repetitions continue to produce an incremental advantage over massed repetitions even if the advantage becomes difficult to measure.

2.1.4. Attenuated spacing effect with re-learning

In most memory experiments, items are studied during one or at most two learning sessions. This deviates from most real-world scenarios where material is reviewed and re-learned on multiple occasions. To create greater ecological validity, Rawson and Dunlosky (2013) taught participants a set of concepts and provided review during five sessions on separate days. Learning during the first session was either massed or spaced. Spacing during the first session had a substantial impact on recall at the start of the second session, but the impact was reduced when the material was restudied and re-learned. Additional studies have confirmed that successive re-learning gradually attenuates effects of initial learning conditions such as spacing and criterion level (Rawson, Vaughn, Walsh,
This has been called the re-learning override effect. Although re-learning override may not seem surprising in hindsight, it is not directly predicted by any theory of the spacing effect. Depending on how the theories are interpreted, they can be used to predict additive, subadditive, or superadditive effects of initial spacing and re-learning (for a review, see Rawson & Dunlosky, 2013). This underscores the difficulty of making quantitative, and sometimes even qualitative, predictions for novel experiment designs based strictly on verbal theories of the spacing effect. The explicit detail and precision required of formal computational and mathematical models may lead to more precise, and hence testable, predictions.

2.1.5. Superadditive learning gains from repetition

How much does repetition improve memory? One baseline is the level of performance that would be expected if the contribution of each item repetition were independent. Based on stimulus sampling theory (Estes, 1955), if the proportion of an item’s available elements encoded during one presentation equals \( x \), the maximum proportion of elements encoded after two presentations equals \( x + x (1 - x) \). Under the assumption that recall probability is linearly related to the proportion of encoded elements (Estes, 1955), the additive baseline can be calculated from observed recall of once-presented items (Begg & Green, 1988; Benjamin & Tullis, 2010). This baseline can be used to evaluate recall of repeated items. The effect of repetition is subadditive if memory for a repeated item falls below the baseline, and superadditive if memory exceeds the baseline.

Begg and Green (1988) conducted five experiments to compare memory for once- and twice-studied items. Cued-recall performance of repeated words was consistently superadditive. In a meta-analysis of free- and cued-recall studies, Benjamin and Tullis (2010) found that performance was subadditive when the time between item repetitions was short, and superadditive when it was long. Somewhat complicating matters, recognition, unlike recollection, is often additive or subadditive (Begg & Green, 1988; Glenberg, 1976; Ross & Landauer, 1978; but see Kahana & Greene, 1993; Russo & Mammarella, 2002 for examples of superadditivity in recognition memory). In some studies that used recognition tests, the spacing interval may not have been long enough to maximize final memory performance and longer intervals may have produced superadditivity. An alternate possibility is that the processes involved in recognition differ in some ways from those involved in recall (Begg & Green, 1988).

An ancillary prediction of stimulus sampling theory, and of the contextual variability hypothesis more generally, is that the greater the elapsed time between any two nonrepeating items, the greater the probability of recalling at least one of the items. This follows from the fact that the two items, inclusively, will be associated with a more diverse set of contextual elements, some of which will overlap with context at the time of test. In a meta-analysis of six studies, Lohnas, Polyn, and Kahana (2011) found that scores based on the probability of retrieving at least one of two nonrepeating items increased by 1%–5% with the amount of space between their presentations. As they note, however, the impact of elapsed time between presentations of different items on retrieval is
substantially smaller than the impact of elapsed time between presentations of the same item, indicating that the spacing effect depends on more than just contextual variability.

2.2. Practical evaluation

Models in cognitive science are typically evaluated in terms of their ability to account for empirical phenomena from well-controlled laboratory studies. These evaluations do not directly speak to a model’s application potential. Accounting for the psychological processes engaged in laboratory studies is necessary but not sufficient. To identify applied criteria for evaluating computational models of the spacing effect, we surveyed papers concerned with applying computational cognitive models to education (Anderson, 2002), fatigue monitoring (Gunzelmann, Veksler, Walsh, & Gluck, 2015), and system design (Gray, 2008). Furthermore, we identified considerations raised in literature from medical (Perez et al., 2013; Stefanidis, Korndorffer, Markley, Sierra, & Scott, 2006) and military training (Jastrzembski, Portrey, Schreiber, & Gluck, 2013; Wisher, Sabol, & Ellis, 1999). From these sources, a set of five considerations consistently emerged (Table 2). These considerations are not unique to models of the spacing effect—they are prerequisites for any computational model to be applied in real-world settings.

2.2.1. Account for effects of training variables on learning and retention

The first applied criterion relates to the breadth of a model’s theoretic adequacy. This is in contrast to account in great detail for one or a small number of findings. Some psychology models focus only on learning (e.g., reinforcement learning; Walsh & Anderson, 2014) or forgetting (e.g., mathematical forgetting functions; Rubin & Wenzel, 1996). Yet these are ongoing, dynamic, competing processes. For example, initial training is typically followed by refresher training to maintain and restore proficiency. To account for knowledge acquisition and retention in such cases, it is necessary to model (1) initial learning, (2) forgetting, and (3) re-learning. Various training factors have different and conflicting effects on knowledge acquisition and retention (Schmidt & Bjork, 1992). Most basically, distributing practice slows acquisition but enhances retention. To evaluate the costs and benefits of different interventions across the complete training lifecycle, computational models must be capable of representing these tradeoffs.

An applied model of memory performance must capture the known effects of multiple scheduling variables on learning and retention. These include at least (a) amount of practice, (b) spacing between practice, (c) number of re-learning sessions, (d) spacing between...

Table 2
Applied considerations for evaluating computational models of the spacing effect

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Account for effects of training variables on learning and retention</td>
</tr>
<tr>
<td>2</td>
<td>Operate on timescales relevant to education and training</td>
</tr>
<tr>
<td>3</td>
<td>Make precise predictions and valid prescriptions</td>
</tr>
<tr>
<td>4</td>
<td>Applicable to a variety of tasks and performance measures</td>
</tr>
<tr>
<td>5</td>
<td>Tractable computational run time</td>
</tr>
</tbody>
</table>
re-learning sessions, (e) length of retention interval, and (f) individual differences in learning and retention. In other words, the first requirement for an applied model of memory performance is to replicate the theoretic criteria documented in the previous section. This list is not exhaustive and could be expanded to include other non-scheduling factors such as difficulty of subject matter, quality and modality of instruction, generalization to related material, processing effects, and transfer from training simulator to real-world setting.

2.2.2. Operate on timescales relevant to education and training

Most spacing effect studies involve retention intervals ranging from seconds to minutes. Fewer involve retention intervals longer than 1 day, and almost none involve retention intervals longer than 1 month (for exceptions, see Cepeda et al., 2006). Likewise, many spacing effect theories account for performance within a single experiment and do not generalize to longer timescales (Bahrick, Bahrick, Bahrick, & Bahrick, 1993). This limits understanding about learning and retention that occurs on an educationally relevant timescale spanning months and years. To be of practical use, models of learning and retention must scale to these longer durations (Anderson, 2002).

Models must continue to account for the moment-by-moment dynamics of memory, however. These dynamics can be leveraged to enhance training and improve performance. For example, the medical community uses just-in-time-refresher training based on the fact that skill decay is inevitable, but that proficiency can be boosted by providing practice immediately before a skill is applied (Barnes, 1998). Anticipating performance in this case requires understanding the interplay between the gradual decay of skill over weeks and months and the transient improvement immediately following practice. Relatedly, re-learning items in sessions separated by multiple days is a key to durable long-term retention. Yet the manner in which items are spaced within a session can greatly affect the number of repetitions needed to achieve mastery within the session, and hence the total time to complete the session (Rawson & Dunlosky, 2013).

2.2.3. Make precise predictions and valid prescriptions

The words fit and prediction, though distinct, are often used interchangeably. A model should account for existing data (i.e., fit), and it should predict behavior in the future or in different circumstances (i.e., prediction). Ideally, predictions and prescriptions should be precise in the sense of being quantitative and testable. The predominant approach in psychology is to gather data from one or a small number of experiments and to fit a model to those data. This involves estimating the values of model parameters that maximize the correspondence between the model’s output and the observed behavior (Shiffrin, Lee, Kim, & Wagenmakers, 2008). However, good fit does not necessarily equate with a persuasive model (Roberts & Pashler, 2000) or with accurate predictions. A model with many degrees of freedom may capture in-sample data, but overfitting such a model may reduce the validity of out-of-sample predictions (Geman, Bienenstock, & Doursat, 1992).
The ability to fit existing data, though important in theoretic evaluation, is not sufficient for applying a model of learning and memory to education or training. The model must also have a \textit{predictive} capability. For example, using data gathered from one or a small number of study schedules, the model must predict acquisition and retention of other novel schedules; that is, out-of-sample schedule prediction. This capability allows exploration of the scheduling design space to quantify and compare the suitability of different schedules (Khajah et al., 2014). Additionally, using data gathered on one timescale, the model must predict retention over longer timescales; that is, out-of-sample temporal prediction. This capability allows the model to be calibrated with training data and used to predict retention. If a model can \textit{predict} performance, the model can be used to \textit{prescribe} education and training events.

Ideally, models of learning and memory should be predictive at the level of individuals (Lindsey et al., 2014; Pavlik & Anderson, 2008). When a cognitive model is applied to multiple participants, each participant may behave according to a different parameterization of the same basic model (Lee & Webb, 2005). This is because cognitive processes and the psychological parameters that control them, such as speed of skill acquisition and rate of forgetting, vary across individuals (Ackerman, 1987; Wixted & Ebbesen, 1997). Moreover, individuals may differ in their training history and prior experience. To maximize the predictive potential of a computational model of learning and memory, these dispositional and situational factors must be accounted for. Doing so enables \textit{personalized} training prescriptions.

2.2.4. Applicable to a variety of tasks and performance measures

Most early studies of the spacing effect involved memorizing lists or associations between pairs of meaningless items (for a review, see Dempster, 1989). Other studies examined the acquisition of motor skills (Lee & Genovese, 1988). More recently, researchers have turned to complex skills and knowledge, such as language syntax (Bird, 2010), mathematical concepts (Rohrer & Taylor, 2006), reading (Seabrook et al., 2005), and military command and control (Arthur et al., 2010). In nearly all cases, distributed practice facilitates retention.

These tasks engage a range of psychological processes. It is unclear whether one theory should account for the benefits of distributed practice across all domains. For example, declarative and procedural knowledge are associated with different learning algorithms, memory representations, and brain networks (Anderson, 1995). Spacing might affect these forms of learning differently. However, the robustness of the spacing effect suggests that a general model, perhaps parameterized uniquely for each domain, can serve as a useful starting point for approximating the key phenomena. Additionally, complex tasks such as trauma assessment and air traffic control simultaneously engage multiple skills and sources of knowledge (Lee & Anderson, 2001). To be applied in domains like these, a computational model should encompass procedural and declarative knowledge, and how they are combined in task performance.
In addition to being applicable to a variety of tasks, a computational model of the spacing effect must account for a variety of outcome measures. Response accuracy is the most common performance measure in spacing studies. However, response times also decrease with experience (Anderson, Fincham, & Douglass, 1999; Newell & Rosenbloom, 1981) and are modulated by practice spacing (Karpicke & Bauernschmidt, 2011; Kwon, Kwon, & Lee, 2015; Pavlik & Anderson, 2008; Spruit, Band, & Hamming, 2015; Wilkins & Rawson, 2013). Both the accuracy and speed of responding are consequential in real-world tasks. As such, a computational model should predict both.

Perceptual-motor tasks involve still other types of outcome measures. There are many examples from medical domains. For instance, cardiopulmonary resuscitation (CPR) performance is evaluated based on four measures: compression rate, compression depth, release height, and hand placement (Oermann, Kardong-Edgren, Odom-Maryon, & Roberts, 2014). Other measures such as time taken, precision of motion, accuracy of clinical decisions, and patient outcomes are used to assess the surgical performance of medical residents (Moulton et al., 2006). One way to model medical skills is to predict motor outputs directly, and to apply the same clinical scoring algorithms to the model’s output. Alternatively, the model might be fitted to sub-skill and composite scores directly. Qualitative expert-based ratings are a final type of performance measure. However, predicting those ratings with a computational model would require corresponding theories of subjective, qualitative assessment that are currently lacking in the cognitive sciences. Fortunately, the increased prevalence of instrumented training devices has led to greater use of quantitative performance measures instead (Kneebone, 2003). Quantitative measures are more suitable for applying computational models.

2.2.5. Tractable computational run time

A final metric for evaluating models of learning and memory is time complexity—the amount of time taken to run as a function of the length of the input (Sipser, 2012). Despite advances in computing, the time complexity of computational models can be prohibitive. This is critical when training prescriptions must be made in near real time, as when deciding what to review next in a computer-based adaptive tutor (e.g., Lindsey et al., 2014). Even when the model is run offline, time complexity may become problematic if the training history contains many events, or when the set of training schedules to choose from is large.

In computer science, time complexity is estimated by counting the number of operations performed by an algorithm, where an operation takes approximately a fixed amount of time to complete. Time is measured as a function of the input. For computational models of the spacing effect, the primary driver of time complexity is the total number of training events, both historic and to be predicted. A useful classification of complexity is whether model run time is sub-linear (a negatively accelerating function of input length), linear (a non-accelerating function of input length), or polynomial (a positively accelerating function of input length). Ideally, model time complexity should be sub-linear or linear.
3. The predictive performance equation

In the preceding section, we identified five criteria for evaluating the theoretic adequacy of computational models of the spacing effect (Table 1) and five criteria for evaluating their applied potential (Table 2). These provide multiple tests along each of two dimensions for judging existing models, and for guiding the development of new models. Next, we present a new model of the spacing effect called the Predictive Performance Equation or PPE (c.f. Jastrzembski & Gluck, 2009). We then evaluate PPE along with two other computational models of the spacing effect according to the complete set of theoretic and applied criteria.

3.1. Model rationale

The acquisition and retention of knowledge are impacted by multiple factors. Most basically, when material is repeatedly tested or studied, memory performance improves. With practice, information can be retrieved more quickly and accurately. The change in performance approximates a power-law—additional practice produces further improvement, but at a diminishing rate (Newell & Rosenbloom, 1981). This learning function is practically universal, and it applies to the acquisition of declarative knowledge and procedural skill (Ritter, Baxter, Kim, & Srinivasmurthy, 2013). Conversely, memory performance declines after periods of non-use. This also approximates a power-law, with rapid loss occurring initially followed by more-gradual, sustained loss (Anderson & Schooler, 1991; Rubin & Wenzel, 1996; Wixted & Ebbesen, 1997). The precise mathematical functions that best capture learning and forgetting curves are debated (e.g., Heathcote, Brown, & Mewhort, 2000; Rubin & Wenzel, 1996); however, all have trajectories resemble a power-law.

The combined effects of these variables are represented in a multiplicative manner in the General Performance Equation (Anderson, 1995; Anderson & Schunn, 2000):

\[ P_n = A \cdot N^c \cdot T^{-d} \]

\( N \) is the amount of practice, \( T \) is the elapsed time since the practice session occurred, \( c \) is the learning rate, \( d \) is the decay rate, and \( A \) is a scaling parameter. Although this equation is useful as a starting point for a computational model of learning and memory, it neglects all phenomena associated with the spacing effect.

Pavlik and Anderson (2005) developed an activation-based theory of spacing through modifications of the declarative memory calculations in ACT-R. In their model, practice repetitions increase long-term memory strength, but as a decreasing function of an item’s current strength. The greater the activation of the item at the time of study, the higher the decay rates of new instances added to memory. The model we propose merges two ideas: (a) practice and elapsed time multiplicatively impact performance (Anderson & Schunn, 2000); and (b) the temporal distribution of practice affects an item’s decay rate (Pavlik & Anderson, 2005).
3.2. Model implementation

To create a more comprehensive account of the spacing effect, we propose the PPE. In PPE, the effects of practice and elapsed time on activation \( M \) of item \( n \) are multiplicative:

\[
M_n = N^c \cdot T^{-d}
\]  

(2)

\( N \) is the number of practice repetitions, \( T \) is elapsed time in seconds, \( c \) is the learning rate, and \( d \) is the decay rate. The expression resembles the General Performance Equation but differs in two ways. First, practice and elapsed time impact activation \( (M_n) \), which is subsequently converted to performance level. Second, the expression no longer includes the scaling parameter \( A \), which is effectively implemented by a different parameter, \( \tau \), introduced later.

Learning rate \( (c) \) is set to 0.1 in all reported simulations. Fixing learning rate across tasks and participants may seem counterintuitive; however, this constraint is consistent with other computational models of learning and memory in which individual and task differences arise from the rate at which information is lost rather than the rate at which it is acquired (e.g., Pavlik & Anderson, 2005). Because decay occurs throughout practice, decay can be seen as impacting rate of acquisition as well. The other practical reason to constrain learning rate is to avoid an identifiability issue with decay. Because the effects of learning and decay are multiplicative (Eq. 2), changes in learning rate can partially compensate for changes in decay.

In the General Performance Equation, elapsed time is treated as the delay between study and test (Anderson & Schunn, 2000). This becomes underspecified when practice is distributed, in which case there may be multiple, vastly different delays between when study repetitions occur and when retention is measured. In PPE, elapsed time is calculated as the weighted sum of the time since each of the previous study or practice events:

\[
T_n = \sum_{i=1}^{n} w_i \cdot t_i
\]  

(3)

The weight assigned to each event decreases with time,

\[
w_i = \frac{t_i^{-x}}{\sum_{j=1}^{n} t_j^{-x}}
\]  

(4)

The variable \( x \) controls the steepness of weighting, wherein higher values result in larger weights for the more recent practice events. The weights always sum to 1 due to the normalization. In all simulations, \( x \) is set to 0.6. The motivation behind Eqs. 3 and 4 is that the age of items in memory should be skewed toward the most recent presentations, but that study history should not be entirely discarded. Indeed, this is the idea behind optimized learning in ACT-R (Petrov, 2006), where the activation of a chunk is primarily influenced by the elapsed time since it was most recently encountered (see...
also Anderson & Lebiere, 1998). Based on Eqs. 3 and 4, elapsed time ($T_n$) may be greater for a twice-studied item than for a once-studied item, owing to the greater amount of time since the first presentation of the twice-studied item. This could lead to the counterintuitive prediction that a twice-studied item would decay more than a once-studied item, actually producing lower memory performance. In actuality, the impact of the second repetition on the practice term ($N$) and on the decay rate ($d$, described below) yields higher retention for twice-studied items.

The variable $d$ in Eq. 2 accounts for decay. To capture the idea that spaced practice produces more stable knowledge, the complete history of lags—or elapsed times—between successive practice opportunities ($\text{lag}_j$) is used to calculate decay,

$$d_n = b + m \cdot \left( \frac{1}{n - 1} \sum_{j=1}^{n-1} \frac{1}{\text{lag}_j + e} \right)$$

(5)

Decay is a linear function of the average of one over the sum of the natural logarithm of the lags between successive practice repetitions. Euler’s number ($e = 2.7818\ldots$) is included in the inner parenthesis to ensure that the denominator is always greater than or equal to one. The quantity inside the summation approaches zero when lags are long, reducing decay to the asymptotic value determined by $b$. The quantity inside the summation approaches one when 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 ($b$).

Predictive Performance Equation treats performance ($P_n$) as a logistic function of activation ($M_n$),

$$P_n = \frac{1}{1 + \exp \left( \frac{\tau - \text{M}_n}{s} \right) }$$

(6)

This forces otherwise unbound activation values to fall between zero and one. The parameters $s$ and $\tau$ control the slope and intercept of the logistic function. Small values of $s$ increase sensitivity of performance ($P_n$) to changes in activation. Small or negative values of $\tau$, in turn, increase the overall level of performance.

Fig. 2 illustrates the dynamics of Eqs. 1–6 for two sample schedules: a massed schedule in which 19 practice repetitions are separated by 5.25 s intervals, and a spaced schedule in which they are separated by 30 s intervals. In both schedules, a retention test is administered on the 20th trial, 96 s after the final (19th) study trial. Values on the x-axis to the left of the vertical line denote time before the end of training, and the value to the right denotes time since the end of training. Each of the plotted points for the massed and spaced schedule corresponds to the time when a practice event is administered. Items in the massed schedule are presented in rapid succession near the end of training, whereas items in the spaced condition are more spread out. Model quantities calculated for each schedule are based on the same parameter values ($\tau = 0.9$, $s = 0.04$, $b = 0.04$, and $m = 0.08$).
Changes in practice ($N$; middle-left panel) with each study repetition are identical between the two schedules, but these changes are condensed toward the end of training in the case of the massed schedule. Elapsed time ($T$; upper-left panel) is considerably longer for the spaced schedule, owing to the greater duration of the lags between repetitions. This limits the activation ($M_n^i$; lower-left panel) and probability of a correct response ($P$; upper-right panel) for the spaced schedule during the learning phase. Greater spacing also reduces decay ($d$; upper-right panel) for the spaced schedule, however. Over the longer retention interval from the end of study to the retention test, the higher decay rate for the massed schedule reduces the activation and the probability of correct responses for items in that condition.

### 3.3. Model summary

Wickens (1998) identified three constraints to limit the class of potential mathematical retention functions: The function must have a real-valued maximum when the length of the retention interval is zero, it must have a real-valued minimum as the length of the retention interval approaches infinity, and it must decrease continuously with the length of the retention interval. The same constraints limit the class of potential mathematical learning functions: The function must have a real-valued minimum when the amount of practice is zero, it must have a real-valued maximum as the amount of practice approaches infinity, and it must increase continuously with practice.
Predictive Performance Equation instantiates a positive power-law of learning \((N^k)\) and a negative power-law of forgetting \((T^{-d})\). Although the isolated effects of these components are complicated by their multiplication and mapping through the logistic response function, they specify power laws of learning and forgetting in terms of an item’s activation. Power functions violate two of Wickens’s (1998); negative power functions are not defined when the length of the retention interval is zero, and positive power functions are not defined when the amount of practice is infinity. However, once the unbounded activation values are passed through the logistic response function in PPE, the model’s outputs are defined for all retention intervals and amounts of practice.

The major innovation in PPE, which differentiates it from the General Performance Equation (Anderson & Schunn, 2000), is that the temporal distribution of an item’s repetitions determines its decay rate. The idea that distributed practice reduces decay has precedent. For example, an item’s decay rate decreases indirectly with time between repetitions in Pavlik and Anderson’s model of the spacing effect (2005), as described in more detail later. Additionally, according to the study-phase retrieval hypothesis, practice gains increase with retrieval difficulty, which depends in part on elapsed time between repetitions (Benjamin & Tullis, 2010; Bjork, 1994; Hintzman, 2004)).

Predictive Performance Equation relates to the New Theory of Disuse (NTD; Bjork & Bjork, 1992), a more general theory of learning and memory. The NTD distinguishes between an item’s storage strength and its retrieval strength. Storage strength corresponds to how well the item is learned, and retrieval strength corresponds to how easily it can be accessed. The latter quantity, retrieval strength, determines the probability of retrieval success. The NTD postulates that the loss of an item’s retrieval strength slows as a function of its storage strength. In PPE, distributing practice can be seen as increasing an item’s storage strength by reducing its decay rate \((d\) in Eq. 2). The retrieval strength of all items decrease with the length of the retention interval, but the loss of retrieval strength is slowest for items with greatest storage strength (i.e., the lowest decay rate).

### 3.4. Alternate computational models of the spacing effect

Other computational models of the spacing effect have been proposed. Here, we focus on two that are among the most widely studied and used. The first model, proposed by Pavlik and Anderson (2005), extends the activation equations from ACT-R and is henceforth referred to as P&A. The second model is a generalization of the Search of Associative Memory model and is henceforth referred to as SAM (Raaijmakers, 2003). We chose these models because they posit different core mechanisms to capture spacing effects (Table 3), and because they have accessible, tractable implementations that enable comparative analysis. Distinguishing between PPE and these models is critical in terms of advancing a mechanistic account of how spacing impacts knowledge acquisition and retention.

In P&A, declarative knowledge is represented by chunks. Chunks have continuously varying activation values that control their accessibility from memory. The total
activation of a chunk reflects the summed activation of the individual instances of the chunk produced by each study exposure (for complete details of the P&A model, see Pavlik & Anderson, 2005; Walsh et al., in press). The activation of a chunk increases with the number of repetitions and decreases with elapsed time since the repetitions occurred.

P&A has one mechanism, activation-dependent trace decay, that produces the spacing effect. The higher the activation of the chunk at the time of a study repetition, the higher the decay rate of the new instance of that chunk that are added to memory. When practice is spaced, an item’s activation rises slowly, yielding lower memory performance during initial acquisition. Because the item’s activation is low, repetitions are stored with lower decay rates. The low decay rates associated with spaced repetitions maintain memory performance across longer intervals from study to test.

In SAM, information is represented by memory traces that contain item information along with other contextual elements incidentally encoded during study (for complete details of the SAM model, see Raaijmakers, 2003; Walsh et al., in press). When a trace is retrieved from memory, new contextual elements that were not previously present are added to the trace. The accessibility of a trace depends on the retrieval cues that are displayed, and the overlap between contextual elements presents throughout study and at test.

Search of Associative Memory has two mechanisms that produce the spacing effect. First, when an item is initially 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 subsequent presentations, the repetitions are not re-encoded in memory. Items that repeat after longer intervals are less likely to remain in the STS and are thus more likely to be re-encoded during each repetition. Second, if the item has left the STS and is retrieved from the LTS, new contextual elements are added to the item’s trace. These contextual elements fluctuate randomly over time and act as retrieval cues during test. Items that repeat after longer intervals are encoded with a more diverse set of contextual elements. The diversity of contextual elements encoded during study increases the probability that some of those elements will also be present during test. Search of Associative Memory contains an additional mechanism called retrieval-dependent updates that captures other effects related to the spacing of practice. New contextual elements are only added to an existing trace if the trace is successfully retrieved. In other words, the benefits of spacing

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Core mechanisms in computational models of the spacing effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Core Mechanisms</td>
</tr>
<tr>
<td>Predictive Performance Equation</td>
<td>Spacing-dependent decay</td>
</tr>
<tr>
<td>P&amp;A</td>
<td>Activation-dependent decay</td>
</tr>
<tr>
<td>Search of Associative Memory</td>
<td>Deficient processing</td>
</tr>
<tr>
<td></td>
<td>Contextual variability</td>
</tr>
<tr>
<td></td>
<td>Retrieval-dependent updates</td>
</tr>
</tbody>
</table>
are contingent upon the original trace remaining accessible throughout study. As described further, this mechanism is needed to account for the non-monotonic effect of spacing interval on retention.

Predictive Performance Equation and P&A each contain four free parameters, and SAM contains six free parameters (Table 4). All three models contain additional parameters which, in principle, could be allowed to vary but are typically fixed (Appendix S1). The ranges were based on estimates reported in previous evaluations of P&A (Pavlik & Anderson, 2005) and SAM (Raaijmakers, 2003). Ranges for all three models’ parameters were expanded to ensure that estimates were not truncated by the upper and lower boundaries. Complete mathematical details about the simulation methodology and the parameter estimation technique are contained in the Supplementary Material.

4. Evaluating computational models of the spacing effect

Having presented three computational models of the spacing effect, we now evaluate those models using the five theoretic and applied criteria. We began by assessing the models’ theoretic adequacy based on post-hoc fits to 14 existing datasets, many of which have been previously used to evaluate models of the spacing effect (c.f. Mozer et al., 2009; Pavlik & Anderson, 2005; Raaijmakers, 2003). Although we only describe a subset of the studies in detail, information about fits to all datasets are included at the end of the section in Table 5.

Table 4
Model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Performance Equation</td>
<td>(\tau) Logistic function intercept</td>
<td>0.5, 1.5</td>
</tr>
<tr>
<td></td>
<td>(s) Logistic function scalar</td>
<td>0.00, 0.10</td>
</tr>
<tr>
<td></td>
<td>(b) Decay intercept</td>
<td>0.00, 0.20</td>
</tr>
<tr>
<td></td>
<td>(m) Decay slope</td>
<td>0.00, 0.20</td>
</tr>
<tr>
<td>P&amp;A</td>
<td>(\tau) Retrieval threshold</td>
<td>(-1.5, 0.5)</td>
</tr>
<tr>
<td></td>
<td>(s) Retrieval noise scalar</td>
<td>0.00, 0.50</td>
</tr>
<tr>
<td></td>
<td>(b) Decay intercept</td>
<td>0.00, 0.50</td>
</tr>
<tr>
<td></td>
<td>(m) Decay slope</td>
<td>0.00, 0.50</td>
</tr>
<tr>
<td>Search of Associative Memory</td>
<td>(\alpha) Fluctuation parameter</td>
<td>0.00, 0.02</td>
</tr>
<tr>
<td></td>
<td>(s) Fluctuation parameter</td>
<td>0.00, 0.08</td>
</tr>
<tr>
<td></td>
<td>(w) Probability that item enters short-term memory store (STS)</td>
<td>0.50, 1.0</td>
</tr>
<tr>
<td></td>
<td>(b) Inter-item information stored on first study trial</td>
<td>0.15, 1.0</td>
</tr>
<tr>
<td></td>
<td>(b_2) Inter-item information stored on subsequent trials</td>
<td>0.15, 1.0</td>
</tr>
<tr>
<td></td>
<td>(\lambda) Rate of decay from STS</td>
<td>0.00, 0.05</td>
</tr>
</tbody>
</table>
4.1. Theoretic criteria

4.1.1. Role of spacing on retention

Massed practice accelerates acquisition and spaced practice enhances retention. This was demonstrated in an experiment by Bregman (1967), where participants memorized associations between pairs of consonant–vowel–consonant (CVC) triples. They were tested on each pair and received feedback during 16 study trials. Bregman manipulated the average number of trials between study repetitions to create massed, intermediate, and spaced conditions. The average numbers of intervening items in the conditions were 1.75, 4.5, and 10, respectively. Retention in all conditions was tested after 32 intervening items, and again after eight additional intervening items. Participants learned pairs most quickly in the massed condition, but retention was highest in the spaced condition (Fig. 3).

Predictive Performance Equation accounts for the effects of spacing on acquisition and retention. When practice is massed, there is little time for instances to decay. Activation of the CVC pairs remains high, accelerating acquisition. However, the short lags between successive study opportunities increase decay. Over the longer interval from study to test,

---

### Table 5
Fits of Predictive Performance Equation (PPE) to empirical datasets

<table>
<thead>
<tr>
<th>Study</th>
<th>RMSE</th>
<th>(r^2)</th>
<th>Number of points</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study</strong></td>
<td><strong>RMSE</strong></td>
<td><strong>(r^2)</strong></td>
<td><strong>PPE</strong></td>
<td><strong>P&amp;A</strong></td>
</tr>
<tr>
<td>Bahrick (1979)(^1)</td>
<td>9.8</td>
<td>6.4</td>
<td>11.8</td>
<td>0.89</td>
</tr>
<tr>
<td>Begg and Green (1988)(^1)</td>
<td>0.0</td>
<td>0.0</td>
<td>1.8</td>
<td>1.00</td>
</tr>
<tr>
<td>Bregman (1967)(^1)</td>
<td>5.4</td>
<td>6.4</td>
<td>7.3</td>
<td>0.95</td>
</tr>
<tr>
<td>Cepeda et al. (2008)(^2)</td>
<td>4.7</td>
<td>3.8</td>
<td>8.3</td>
<td>0.97</td>
</tr>
<tr>
<td>Cepeda et al. (2009)(^2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td>3.2</td>
<td>3.9</td>
<td>6.5</td>
<td>0.96</td>
</tr>
<tr>
<td>Experiment 2A</td>
<td>4.1</td>
<td>5.3</td>
<td>10.1</td>
<td>0.97</td>
</tr>
<tr>
<td>Experiment 2B</td>
<td>4.1</td>
<td>5.6</td>
<td>8.5</td>
<td>0.98</td>
</tr>
<tr>
<td>Glenberg (1976)(^1)</td>
<td>3.6</td>
<td>4.5</td>
<td>3.7</td>
<td>0.88</td>
</tr>
<tr>
<td>Pavlik and Anderson (2005)(^2)</td>
<td>4.8</td>
<td>5.4</td>
<td>6.4</td>
<td>0.97</td>
</tr>
<tr>
<td>Rawson and Dunlosky (2013)(^2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td>5.6</td>
<td>3.4</td>
<td>7.8</td>
<td>0.88</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>4.8</td>
<td>3.8</td>
<td>7.3</td>
<td>0.91</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>4.4</td>
<td>4.3</td>
<td>6.5</td>
<td>0.94</td>
</tr>
<tr>
<td>Rumelhart (1967)(^1)</td>
<td>1.9</td>
<td>1.8</td>
<td>1.9</td>
<td>0.99</td>
</tr>
<tr>
<td>Young (1971)(^1)</td>
<td>4.0</td>
<td>4.0</td>
<td>2.7</td>
<td>0.83</td>
</tr>
<tr>
<td>Mean</td>
<td>4.4</td>
<td>4.2</td>
<td>6.5</td>
<td>0.93</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2.1</td>
<td>1.7</td>
<td>3.0</td>
<td>0.05</td>
</tr>
</tbody>
</table>

PPE: Predictive Performance Equation; SAM: Search of Associative Memory.

\(^1\)Within-subjects design.

\(^2\)Between-subjects design.
CVC pairs from the massed condition decay more rapidly than instances from the spaced condition, reducing retention.\(^8\)

P&A also accounts for the effects of spacing on acquisition and retention. As in PPE, the activation values of instances remain high across the short intervals separating massed repetitions, accelerating acquisition. However, greater activation causes subsequent instances to be stored with higher decay rates. Over the longer interval from study to test, the higher decay rates associated with massed presentations yield lower retention.\(^9\)

Finally, SAM accounts for the effects of spacing on acquisition and retention. During the acquisition phase, massed items are more likely to remain in the STS across the short intervals between their repetitions. This enables accurate responses but prevents items from being re-encoded in memory, effectively limiting the number of practice repetitions they receive. All items leave the STS during the longer interval from study to test. Because massed items were re-encoded fewer times during the acquisition phase, they are more difficult to retrieve during test trials.\(^10\)

Fig. 3. Performance in massed, intermediate, and spaced conditions during 16 study trials, and on two later retention tests (R1 and R2). Circles show observed performance, and lines show data from Predictive Performance Equation (PPE), P&A, and Search of Associative Memory (SAM) calibrated for best fit to the Bregman (1967) data.
4.1.2. Relationship between retention interval and optimal spacing interval

Increasing the amount of time between study repetitions improves performance to a point, after which it impairs retention. Cepeda et al. (2009) demonstrated a non-monotonic effect of ISI in two experiments. In the first experiment, participants learned obscure trivia facts. During the first session, they viewed facts and were quizzed to a criterion of one correct response per fact. During the second session, they were quizzed again to a criterion of one correct response per fact. Retention was assessed at the start of the third session. The first and second sessions were separated by an ISI ranging from 0 to 14 days, and the second and third sessions were always separated by 10 days. In Experiment 2, participants learned trivia facts as well as the names of obscure objects. The first and second sessions were separated by an ISI ranging from 0 days to 6 months, and the second and third sessions were always separated by 6 months. In both experiments and for both types of stimuli, retention at the start of the second session decreased with the length of the ISI (Fig. 4, left column), reflecting the standard forgetting curve. Retention at the start of third session increased and then decreased with the length of the ISI separating the first and second sessions (Fig. 4, right column). Retention was maximal for intermediate values of the ISI in both experiments and for both stimulus types.

All three models account for the effect of ISI on retention performance at the start of the second session (Fig. 4, left column). In PPE and P&A, increasing elapsed time from Session 1 to Session 2 results in more decay, reducing memory performance. In SAM, increasing elapsed time from Session 1 to Session 2 results in more contextual fluctuation, also reducing memory performance.

Predictive Performance Equation accounts for the non-monotonic effect of ISI on retention at the start of the third session (Fig. 4, left column). Increasing the lag between sessions reduces the decay rate, which improves final retention (Eq. 5). However, increasing the lag also increases the total elapsed time between the third session and the earliest presentations of items in the first session, which impairs retention (Eq. 2). The conflicting effects of ISI on decay rate and elapsed time produce a non-monotonic relationship between ISI and retention.

P&A also accounts for the non-monotonic effect of ISI on final retention (Fig. 4, right). Retention depends on the combined strength of instances added to memory during the first and second sessions. If the ISI is brief, activation of the item will remain high at the start of the second session. Consequently, the decay rates of instances added during the second session will be high, and those instances will be quickly forgotten. If the ISI is long, the total elapsed time between the first and third sessions will be substantial. Consequently, instances from the first session will be forgotten by the final test. To maximize retention, the ISI must be long enough to allow activation to decrease before the second session, but not so long that instances from the first session are completely lost before the test.

Search of Associative Memory weakly accounts for the non-monotonic effect of ISI on final retention (Fig. 4, right). As elapsed time between repetitions increases, contextual fluctuation occurs, and new contextual elements are added to the item’s trace. The greater the elapsed time between repetitions, the greater the number of new contextual elements added to a trace. Contextual elements are only added to an existing trace if the item is
successfully retrieved, however. The conflicting effects of increasing the number of new contextual elements while decreasing the probability of retrieving an item at the start of the second session produces the non-monotonic relationship between ISI and final retention.

Experiments have shown that the ISI that maximizes retention varies with the length of the RI. Cepeda et al. (2008) examined the joint effects of ISI and RI over the course of weeks and months. During the first session of their experiment, participants read trivia

Fig. 4. Retention performance at the start of the second session (left column) and the third session (right column). Data are plotted as a function of the inter-session interval separating the first and second sessions. The third session always occurred 10 days (Experiment 1) or 6 months (Experiment 2) after the second session. Circles show observed performance, and lines show data from Predictive Performance Equation (PPE), P&A, and Search of Associative Memory (SAM) calibrated for best fit to the Cepeda et al. (2009) data.
facts and were quizzed to a criterion of one correct response per fact. During the second
session, they were quizzed again to a criterion of one correct response per fact. Retention
was assessed during a final test session. Cepeda et al. (2008) varied the ISI between
learning sessions in seven levels ranging from 0 to 105 days, and the RI before the final
test in four levels ranging from 7 to 250 days. They partially crossed these factors to cre-
ate a total of 26 ISI-RI conditions. For each ISI, retention decreased with the length of
the RI, replicating the standard forgetting curve (Fig. 5, left). For each RI, performance
increased and then decreased with the length of the ISI, replicating the non-monotonic
effect of ISI on retention. Most important, the ISI that maximized retention varied with
length of the RI and as an increasing function of it.

Predictive Performance Equation captures the interaction between ISI and the RI on
final retention (Fig. 5, top). When the RI is moderate or long, elapsed time at final
test is also long regardless of the length of the ISI. In such cases, increasing the ISI
reduces decay while only slightly increasing elapsed time. Conversely, when the RI is
short, elapsed time at test may be small if the ISI is also short. In such cases,

![Graphs showing retention performance as a function of ISI and RI](image)

Fig. 5. Final retention performance as a function of the ISI and retention interval. Circles show observed per-
formance, and lines show data from Predictive Performance Equation (PPE), P&A, and Search of Associative
Memory (SAM) calibrated for best fit to the Cepeda et al. (2008) data.
increasing the ISI reduces decay rate but has the larger negative effect of increasing elapsed time.

P&A captures the interaction between the ISI and the RI in a similar manner (Fig. 5, middle). When the RI is moderate or long, instances from the first study session will have low strength at test regardless of the length of the ISI. In such cases, increasing the ISI reduces the decay rate of instances encoded during the second study session, enhancing final retention. When the RI is short, instances from the first study session may have moderate strength at test if the ISI is short as well. In such cases, reducing the ISI improves final retention by minimizing elapsed time since the earliest repetitions of an item.

Search of Associative Memory produced a weak interaction between ISI and RI (Fig. 5, bottom). The optimal ISI increased from zero for RIs of 70 days or less, to 21 for the RI of 350 days. This reflects two opposing forces. First, an item must be successfully retrieved in the second session for new contextual elements to be added to its trace. Second, as the RI increases, only items encoded with many contextual elements remain retrievable. When the RI is short, the addition of even a few new contextual elements with successful retrievals during the second session enhances final retention. This shifts the optimal ISI toward shorter values. When the RI is long, only the addition of a large number of new contextual elements in the second session is enough to enhance final retention. This shifts the optimal ISI toward longer values. Thus, while fewer items may be recallable after a long ISI, such a delay is necessary for any items to remain recallable after a long RI.

Using PPE’s best fitting parameter estimates from the Cepeda et al. (2008) data, we extended these results by simulating retention for all combinations of ISIs ranging from 0 to 100 days and RIs ranging from 0 to 350 days. The resulting surface (Fig. 6, left) contains a “temporal ridgeline of optimal retention” similar to the one reported by Cepeda et al. (2008); the optimal ISI increases proportionally with the RI (Fig. 6, right). The slope of the function relating the optimal ISI to the RI is <1, and the ratio of the optimal ISI to the RI decreases with the length of the RI. Also consistent with Cepeda et al. (2008), performance drops sharply for durations shorter than the optimal ISI, but more gradually for durations longer than the optimal ISI (Fig. 6 left).

Two observations warrant further comment. First, PPE appears to predict errorless memory when the RI is short. This is only true for the shortest RI shown (0 days). Performance drops by 10% by the second RI plotted in Fig. 6 (3 days). Second, Cepeda et al. (2008) fit a mathematical function to retention data from their study and generated a retention surface like the one in Fig. 6. The results of their analysis were consistent with PPE, with the exception that the mathematical function they fitted predicts that the optimal ISI will decrease as a power function of the length of the RI (c.f., Mozer et al., 2009), whereas PPE (and P&A) predict that the optimal ISI will decrease as a linear function of the length of the RI. Additional studies are needed to resolve the exact shape of the function. These studies must include longer RIs, where power and linear functions diverge to a greater extent.
4.1.3. Increased benefit of spacing with amount of practice

Pavlik and Anderson (2005) examined the effects of varying number of practice repetitions and spacing. In their experiment, participants learned Japanese–English word pairs. Pavlik and Anderson varied the number of times that each word pair appeared (1, 2, 4, or 8), and the number of trials between repetitions (2, 14, or 98). They tested retention during a second session that occurred 1 or 7 days later. During the learning session (Session 1), performance increased with each repetition, and decreased with the spacing between repetitions (Fig. 7, left). During the test session (Session 2), retention was greatest for items with the most repetitions and the longest spacing during the earlier learning session (Fig. 7, right). Most important, the size of the spacing effect in Session 2 increased with the number of repetitions from Session 1.

Predictive Performance Equation accounts for the interaction between amount of practice and spacing. Because the lag before the first presentation of a word pair is infinity (lag1 in Eq. 5), decay for all presentation schedules initially equals the intercept parameter b. The sequence of lags between subsequent presentations move decay from this starting value to distinct values for each study schedule. This enhances the spacing effect. Additionally, item learning and decay are multiplied in Eq. 2. Consequently, the effects of spacing on item decay are scaled by item learning, which increases with number of practice repetitions.

P&A also accounts for the interaction between amount of practice and spacing. The activation values of all word pairs are initially equal. As spacing gives rise to differences in the activation values of massed versus spaced items, the decay rates of new instances
added to memory begin to diverge. This effect is magnified by the number of repetitions, each of which contributes incrementally more to retention in the spaced condition than in the massed condition.

Fig. 7. Learning (Session 1) and retention (Session 2) as a function of the number of repetitions and the spacing between repetitions. Circles show observed performance, lines show Predictive Performance Equation (PPE), P&A, and Search of Associative Memory (SAM) calibrated for best fit to the Pavlik and Anderson (2005) data.
Finally, SAM accounts for the interaction between amount of practice and spacing. Increasing the number of repetitions only enhances retention if an item leaves the STS and is re-encoded. This occurs when items are separated by 14 or 98 trials, but not when they are separated by two trials. In fitting SAM to these data, Pavlik and Anderson (2005) relaxed the retrieval-dependent update assumption and allowed contextual elements to be added to traces regardless of whether or not the retrieval attempt was successful. When the retrieval-dependent update assumption is included, as in the current simulations, the benefits of encoding a more diverse set of contextual elements in the 98-trial spacing condition are offset by the greater number of retrieval failures that occur in that condition.

4.1.4. Attenuated spacing effect with re-learning

Rawson and Dunlosky (2013) examined whether the benefits of spacing persist across re-learning by teaching participants psychology terms and administering multiple re-learning sessions. On Day 1, they defined the terms and tested participants to a criterion of 1 or 3 correct responses per term. On Days 3 and 8, they re-tested participants to a criterion of 1 correct response per term. Finally, on Day 10, they re-tested participants once. Rawson and Dunlosky varied the number of intervening items between repetitions on Day 1 (0, 1, 3, or 7 intervening items), and they set the number of intervening items to 7 on all subsequent days. The benefit of spacing, which was substantial on the initial retention test at the start of Day 3, diminished with each re-learning session (Fig. 8).

Predictive Performance Equation accounts for the diminishing benefit of initial spacing with subsequent re-learning. This occurs because the introduction of lags between days gradually makes the schedules more similar to one another, offsetting the Day 1 manipulation. P&A also accounts for the diminishing benefit of initial spacing with subsequent re-learning. As in PPE, the introduction of lags between days gradually overwhelms the effects of the manipulation administered on Day 1. Additionally, when practice is spaced on Day 1, instances of the corresponding terms are stored with low decay. Activation of those terms is higher at the onset of Day 3, as reflected by the greater accuracy during the retention test administered at the start of the session. Because spaced terms have high activation at the start of Day 3, new instances added to declarative memory have higher decay rates relative to terms that were massed on Day 1. Paradoxically, by increasing retention at the start of Day 3, spaced practice reduces the benefits of subsequent encodings on that day and on later days.

Finally, SAM accounts for the diminishing benefit of initial spacing with subsequent re-learning. When an item is repeated, only the contextual elements not previously encoded are added to the trace. When practice is spaced on Day 1, more contextual elements are added to an item’s traces and, consequently, fewer new elements remain on subsequent days. When practice is massed on Day 1, the greater number of contextual elements encountered and encoded at the start of Day 3 offset the earlier spacing manipulation.
4.1.5. Superadditive learning gains from repetition

To test whether the effects of item repetition are subadditive, additive, or superadditive, Begg and Green (1988) taught participants noun pairs. They compared participants’ recall of twice-presented pairs with their recall of two separate once-presented pairs. To control for serial position effects, they used a yoked between-subjects design; they placed once-presented pairs in the exact positions occupied by the twice-presented pairs from other participants. Begg and Green found that memory for the repeated pairs exceeded the independence baseline calculated from the two once-presented pairs (Fig. 9). A later meta-analysis supported these results by showing superadditivity in many cued-recall

---

Fig. 8. Percent correct on the first re-test during each day. Bars show observed performance and circles show data from Predictive Performance Equation (PPE), P&A, and Search of Associative Memory (SAM) calibrated for best fit to the Rawson and Dunlosky (2013) data.
experiments (Benjamin & Tullis, 2010). Superadditivity was most consistently reported in experiments with long ISIs.

Predictive Performance Equation accounts for the superadditive effect of repetition (Fig. 9, left). Although the learning function in PPE is a subadditive power law (Eq. 2), the logistic response function (Eq. 6) is non-linear. Consequently, two items studied once may both have a low probability of recall, whereas one item studied twice may have a far greater probability of recall. Predictive Performance Equation also predicts a subadditive effect of repetition at short ISIs. When the ISI is brief, the second presentation of an item increases the decay rate. Greater decay may cause the probability of retrieving the twice-studied pair to fall below the joint probability of retrieving at least one of the once-studied pairs.

P&A and SAM also account for the superadditive effect of repetition (Fig. 9 center and right). In P&A, item strength is summed across instances. Although the strengths of the two instances are added, the logistic response function (Eq. 2) is non-linear. Consequently, two instances, which alone have a low probability of recall, may yield a far greater probability of recall when combined. Relatedly, retrieval probability in SAM is a non-linear function of number of encoded contextual elements. Consequently, the probability of retrieving a trace with many contextual elements may exceed the probability of retrieving at least one of two traces with fewer contextual elements.

Benjamin and Tullis (2010) evaluated superadditivity in studies of free- and cued-recall as a function of ISI alone because there were too few studies to consider ISI and RI jointly. In the absence of adequate existing data, computational simulations can be used to enumerate the prediction space. To examine how ISI and RI jointly modulate the impact of item repetition, we used PPE, parameterized based on Begg and Green (1988),

![Fig. 9. Observed recall of once-presented items (blue and red circles) and twice-presented items (black circle) with the ISI used by Begg and Green (1988). Green circle shows prediction of independence baseline. Curves show data from Predictive Performance Equation (PPE), P&A, and Search of Associative Memory (SAM) calibrated for best fit to the Begg and Green data.](image-url)
to test for superadditivity for all combinations of ISIs ranging from 1 to 500 s and RIs ranging from 1 to 6,000 s. Fig. 10 (right) shows the shortest ISI to produce superadditivity for each RI. The red star corresponds to the RI used by Begg and Green (1988). For each RI, all ISIs greater than the plotted point also produce superadditivity. These results indicate that superadditivity is most likely to be found when the ISI is sufficiently long, and when the RI is also long. The right-skewed, inverted U-shape relates to the temporal ridgeline of optimal retention reported by Cepeda et al. (2008). As the RI increases, values of the ISI that facilitate final retention increase as well. Values of the ISI that facilitate final retention, in turn, are most likely to produce superadditivity for twice-studied items.14

4.1.6. Theoretic criteria: Summary

Table 5 contains metrics of model fits to 14 datasets, including the eight reported in the previous section. These results support two clear conclusions. First, all models accounted qualitatively for the complete set of spacing-related phenomena. In this sense, they satisfy the theoretic criteria outlined in Table 2. Second, notwithstanding this success, PPE and P&A consistently provided a better quantitative fit to the experimental data than SAM did. This may relate to the fact that several data sets involved spacing and retention intervals spanning multiple days. Search of Associative Memory may be less suitable for modeling performance across such long intervals. Indeed, average RMSE (based on percentage correct ranging from 0 to 100) across the five experiments that used a single-session design was similar for the three models (PPE = 3.0, P&A = 3.4,

Fig. 10. Predictive Performance Equations predictions of ISIs that produce superadditive effects of repetition for retention intervals (RIs) of varying durations. Red point corresponds to RI used by Begg and Green (1988).
SAM = 3.5), whereas RMSE across the nine multi-session experiments was about twice as large for SAM (8.1) versus PPE (5.2) and P&A (4.7).

Retrieval-dependent updates allow SAM to account for the non-monotonic effect of spacing interval on retention (Raaijmakers, 2003) and are thus an essential component of the model. Yet the assumption that a trace is only updated if the retrieval is successful, and that the trace is otherwise lost, may be too strong. Retrieval failures are inevitable when items are separated by long inter-trial and inter-session intervals, and yet people benefit from such distributed practice. Additional work is needed to determine whether and how to modify the retrieval-dependent update assumption to generalize SAM to longer, educationally relevant timescales.

Are group-averaged data from each study sufficient to estimate parameters for PPE and P&A, and do these data provide a strong test of the models? This is certainly a concern for the study by Begg and Green (1988), which only contained three data points. However, the number of points in the other studies ranged from 10 to 64 (Table 5). In several of the experiments, these points correspond to performance across two or more phases of learning. For example, five of the simulation studies involved learning and retention (Bregman, 1967; Cepeda et al., 2009, Young, 1971), and one involved re-learning as well (Pavlik & Anderson, 2005). Such experiments provide an exceptionally strong test of the models. When a model is fitted to data from one learning phase (i.e., initial learning or final retention), variations in parameter estimates may obscure model inadequacies. The parameter values that best capture retention may vary widely from the values that best capture initial acquisition or re-learning. When one set of parameter values is estimated to simultaneously account for initial acquisition, retention, and re-learning, one can more directly assess whether the model’s mechanisms are sufficient to account for the impact of an experimental manipulation, such as spacing, on performance across all phases of learning.

4.2. Applied criteria

The previous set of simulations addresses the theoretic adequacy of PPE, P&A, and SAM. Yet the question remains, can these models actually be used to inform training and education decisions? Next, we evaluate the models’ application potential. The goal of this evaluation is not to reject one or more of the models. Rather, the goal is to identify potential barriers to apply each of the models. These barriers must be identified to guide model development, algorithm development, and hardware development, and to facilitate technology transition.

4.2.1. Account for effects of training variables on learning and retention

Predictive Performance Equation is a model of how practice scheduling impacts learning and retention. Increasing practice ($N$) produces learning, and increasing elapsed time ($T$) produces forgetting (Eq. 2). Most basically, these mechanisms allow PPE to model initial learning, forgetting, and re-learning. Furthermore, the spacing term in PPE captures effects of training variables related to the temporal distribution of practice (i.e., amount
of practice, spacing between practice, number of re-learning sessions, and spacing between re-learning sessions). The ability of PPE to account for the range of theoretic phenomena reported in the previous section demonstrates that it satisfies this criterion. Predictive Performance Equation is not unique in this regard. P&A and SAM include different mechanisms that also allow them to capture effects of these training variables on learning and retention.

To be of greatest use in an applied context, a computational model should account for other variables besides amount and timing of practice. All three models currently lack a theory of how individual differences in aptitude, personality, and prior knowledge affect performance (Ackerman, 1987; Landsberg et al., 2012). Given a reasonable amount of data, parameters could be estimated separately for each individual to represent these differences. The models also lack a theory of how acquisition and retention vary by stimulus and task type (Arthur, Bennett, Stanush, & McNelly, 1998; Reber, 1989), mode of instruction (Kalyuga, Chandler, & Sweller, 2000), or quality of training experience (Rudolph, Simon, & Raemer, 2007). Finally, the models are agnostic with respect to the nature of the learning event. For example, none explicitly represent the benefits associated with active retrieval versus passive study (i.e., the testing effect; Roediger & Karpicke, 2006). Parameters could be estimated separately based on features of the task, instruction, and event types (c.f. Pavlik & Anderson, 2008; Sense, Behrens, Meijer, & Rijn, 2016), but ideally the models would contain mechanisms to account for the effects of such variables.

4.2.2. Operate on an educationally relevant timescale

In the previous sections, PPE and P&A were shown to account for data from studies that take place on timescales ranging from seconds to months. These cover the short-to-medium range of educationally relevant times. At the other extreme, Bahrick (1979; Bahrick & Phelps, 1987) taught participants Spanish–English vocabulary pairs and tested retention years later. After an initial learning session, participants completed two or five additional sessions. Each session began with a test, and pairs that were not correctly recalled were reviewed. Sessions were evenly spaced at intervals of 0, 1, or 30 days. Final retention was assessed after 30 days, and in some cases again after 8 years. Figure 11 shows data from the study, along with the fits of PPE, P&A, and SAM. Participants showed greatest retention when study sessions were spaced by 30 days, followed by 1 day, and then by 0 days. All three models account for the acquisition and retention of Spanish vocabulary over 8 years, although SAM’s predictions are virtually identical for the 0 and 1 Day ISI conditions.

Predictive Performance Equation captures spacing effects over timescales ranging from seconds to months. The impact of spacing in PPE is modulated by the logarithm of the lags between repetitions (Eq. 5). This ensures high sensitivity to small variations in spacing over short intervals such as those within experiment sessions, and continued but diminishing sensitivity to variations in spacing over longer intervals such as those between experiment sessions.
Pavlik and Anderson (2005) noted that ACT-R tends to predict too much forgetting across weeks and months. To deal with this, they multiplied elapsed time between sessions by a scaling parameter \((h = 0.025)\). They supposed that there was less destructive interference from events that occurred between sessions than within (Anderson et al., 1999). In effect, this reduces decay during the time between sessions. The model results in Fig. 11 are based on simulations that do not include psychological scaling of time between sessions. Somewhat surprisingly, scaling was not needed to account for these results or for results of any of the multi-session studies in Table 5. The estimates for the model’s decay parameters were systematically smaller in multi-session experiments than in single-session experiments, however, indicating compensation for the longer temporal intervals. One potential advantage to include something like psychological scaling in P&A is to reduce differences in parameter estimates between single- and multi-session experiments.

Search of Associative Memory could also be fitted to the data. The estimate for the contextual fluctuation parameter was extremely low to allow performance to remain above floor after the long intervals separating sessions. The estimate was so low that the probability of retrieval in the 0 and 1 Day ISI conditions equaled the probability that an item had entered the LTS during any of the previous trials. Looking beyond this single data set, the estimates for the model’s contextual fluctuation parameter were systematically smaller in multi-session experiments than in single-session experiments (Table 5). A shared limitation of P&A and SAM, then, is that certain model parameters bound the timescale over which the models operate.

4.2.3. Make precise predictions and valid prescriptions

To be used to prescribe training, a model must accurately predict performance for novel schedules (i.e., out-of-sample schedule prediction), and it must predict future performance (i.e., out-of-sample temporal prediction). A series of experiments by Rawson and Dunlosky (2013) provide an opportunity to test out-of-sample schedule prediction. Rawson and Dunlosky conducted three experiments (one of which was described in Section 4). In each experiment, participants learned psychology terms during an initial session. They were then quizzed to a criterion of one correct response per term during multiple sessions delivered on subsequent days. In Experiment 1, items were presented at lags of 0, 1, 3, and 7 on Day 1, and in Experiment 2, items were presented at lags of 0 and 7 on Day 1. All items were presented at lag 7 during every subsequent session. Experiment 2 included an extra session given 38 days after the final test from Experiment 1. Experiment 3 differed from Experiments 1 and 2 in that spacing was manipulated during the initial session and during all subsequent sessions. Items were presented at lags of 0 or 7 on Day 1, and at lags of 0 or 7 on all subsequent days. Initial spacing and subsequent spacing were crossed, creating four unique conditions in Experiment 3.\(^{15}\)

To evaluate whether PPE, P&A, and SAM make accurate out-of-sample schedule predictions, we fit the models to data from one experiment and used the fitted models to predict data from the other two experiments. We expected that parameters would generalize between Experiments 1 and 2 because the experiments were highly similar—both
manipulated spacing within the first session only. We were unsure of whether parameters would generalize from Experiments 1 and 2 to Experiment 3, however, because spacing was manipulated during all sessions in Experiment 3.

Fig. 11. Acquisition of Spanish–English word pairs across three or five sessions, and percent correct retrieval after 1 month (R1) and 8 years (R2). Circles show observations, and curves show data from Predictive Performance Equation (PPE), P&A, and Search of Associative Memory (SAM) calibrated for best fit to the Bahrick (1979) data.
Fig. 12 shows the RMSE (based on percentage correct ranging from 0 to 100) when using best fitting model parameters from the experiment listed in a column to predict data from the experiment listed in the row. Colors denote the increase in RMSE for out-of-sample prediction relative to fitting. For all models, parameters generalized between Experiments 1 and 2 as indicated by the small increase in RMSE for out-of-sample prediction. For PPE and SAM, parameters also generalized between Experiments 1 and 3, and between Experiments 2 and 3. The mean increase in RMSE for out-of-sample prediction was 4.1 for PPE and 4.6 for SAM. The mean increase in RMSE for out-of-sample prediction was somewhat greater for P&A, however (6.3).

The series of experiments by Rawson and Dunlosky also provides an opportunity to test out-of-sample temporal prediction. In particular, Experiment 2 contains inter-session intervals ranging from 2 to 28 days. We asked, given data from sessions 1 to \(N - 1\), how well would the models predict retention at the start of session \(N\) (i.e., accumulative one-step look-ahead; Wagenmakers, Grunwald, & Steyvers, 2006)?

Table 6 shows the RMSE (based on percentage correct ranging from 0 to 100) for accumulative one-step look-ahead prediction. For all models, predictions for Days 8 and 10 are reasonably accurate. Performance spanning the 2-day ISI between the initial learning session on Day 1 and the first re-learning session on Day 3 is sufficient for estimating parameter values that generalize across the longer ISI between Days 3 and 8. PPE and P&A also make reasonably accurate predictions for Day 38, whereas SAM is less accurate. SAM’s contextual fluctuation parameters, estimated from ISIs ranging from 2 to 5 days, generalize poorly to the longer ISI separating Days 10 and 38.

To be maximally useful, a computational model of the spacing effect should make individual-level out-of-sample temporal predictions. Such predictions, if valid, could be used to prescribe personalized review (Mettler et al., 2016; Pavlik & Anderson, 2008). We have begun to explore this issue in PPE and have found that the model can, in fact, make accurate individualized predictions for a range of skill-based and declarative tasks (Jastrzembski et al., 2017). However, all three models should be evaluated further based on their ability to predict individual performance and prescribe personalized review.
4.2.4. Applicable to a variety of tasks and performance measures

Predictive Performance Equation is an abstract mathematical model of performance. Its outputs, which fall between 0 and 1, can be scaled to any bounded performance measure. For instance, in all simulations from Section 4.1, PPE’s output was treated as the probability of a correct response. PPE can be applied to tasks that involve declarative knowledge, procedural knowledge, or combinations of both. However, further research is needed to determine whether PPE accounts equally well for performance in all of these cases, and whether parameter estimates systematically vary among tasks that involve different combinations of declarative and procedural knowledge.

P&A is based on ACT-R, an integrated cognitive architecture. ACT-R predicts multiple performance measures, including response accuracy, response time, and neural activation (Anderson et al., 2004). By extension, P&A also predicts these performance measures. In past applications of P&A, response latencies have been used to calibrate model parameters when participants commit a small number of retrieval errors (Sense, Behrens, Meijer, & van Rijn, 2016). Additionally, P&A has been used to predict retrieval latencies to identify study schedules that will allow students to complete the greatest number of review trials in the least amount of time (Pavlik & Anderson, 2008; see also Sense et al., 2016). ACT-R distinguishes between declarative and procedural knowledge, and P&A is specifically a model of declarative knowledge. An open question is how to extend P&A to account for the acquisition and retention of procedural and hybrid skills.

Search of Associative Memory was developed to account for the probability of recalling declarative knowledge in a range of memory paradigms. It is unclear whether SAM can be applied to the acquisition and retention of procedural or hybrid skills, nor was it intended to. It is also unclear whether SAM can be extended to performance measures besides percentage of recall or recognition, such as response time.

4.2.5. Tractable computational run time

Fig. 13 shows the amount of time for PPE, P&A, and SAM to predict performance at a future point following \( n \) training events. For example, when \( n = 100 \), the models are given a training history of 100 events and they predict performance for event \( n + 1 \). Run time following 100 training events remains low for PPE (~1 ms) and moderate for P&A (~125 ms). P&A must iteratively calculate item activation from events 1 through \( n \) to determine the instance-specific decay rates that give rise to the prediction for event

<table>
<thead>
<tr>
<th>Retention Interval</th>
<th>Day 8</th>
<th>Day 10</th>
<th>Day 38</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-step look-ahead</td>
<td>Predictive Performance Equation</td>
<td>4.0</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>P&amp;A</td>
<td>8.3</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>Search of Associative Memory</td>
<td>7.5</td>
<td>10.7</td>
</tr>
</tbody>
</table>
The iterative calculation of activation following events \( I \) through \( n \) gives P&A subquadratic time complexity. All of PPE’s calculations are based on elapsed time between adjacent repetitions, and elapsed time between event \( n + 1 \) and all previous events. This gives PPE linear time complexity.

Run time for SAM exceeds 10 s after 50 training events. The reason that SAM is so computationally expensive is that the prediction on trial \( n + 1 \) must be conditioned on a probabilistic sequence of model outcomes that may have occurred during the previous \( n \) trials. These include the first trial when the item was successfully encoded in the STS, the first trial when the item was no longer available from the STS, and the history of retrieval successes and failures once the item left the STS. Considering only the third source of probabilistic model outcomes, the history of retrieval successes and failures from the LTS, \( 2^n \) possible histories exist for an item that entered and left the STS on the first trial. Although many trial histories can be pruned, the set of histories grows exponentially with the number of events, giving SAM superpolynomial time complexity.

Although the cost of a single model run is minor, practical applications may require hundreds or thousands of runs. For example, the first step in personalized training prescription is to fit the model to an individual’s calibration data, which involves enumerating model predictions over many combinations of parameter values to find the best fitting set. Model fits in Table 5 are typically based on 100 iterations of a simplex algorithm (Nelder & Mead, 1965), with five parameter sets tested per iteration (number of free parameters plus one), for a total of 500 model runs per individual. The second step is to evaluate model performance for a large number of (as of yet untested) candidate...
schedules in order to identify the best schedule. As the number of training events or sessions to simultaneously plan for increases, a combinatorial explosion occurs. For instance, over a 30-day training period, two sessions can be assigned to distinct days in 870 different ways, three sessions can be assigned in 24,360 ways, and four sessions can be assigned in 657,720 ways. Ultimately, techniques for adaptively exploring the model parameter space and the schedule parameter space are necessary, but a first defense against the combinatorial explosion is to minimize a model’s time complexity.

4.2.6. Applied criteria: Summary

Predictive Performance Equation satisfies several criteria for an applied model of the spacing effect. It accounts for effects of training variables on learning and retention, it operates across multiple timescales, it is capable of out-of-sample schedule and temporal prediction, and it has computationally tractable run time. Notwithstanding these strengths, PPE currently lacks a theory of how individual differences impact performance—these differences are partially captured by variations in the model’s parameter estimates. Predictive Performance Equation also does not represent variables besides amount and timing of training. Finally, PPE is somewhat limited in that it predicts bounded performance measures only.

The applied criteria evaluations for P&A and SAM were somewhat more mixed. The strength of P&A relative to PPE is that P&A is applicable to a greater range of measures, including retrieval latencies and neural activation. P&A performed more poorly for out-of-sample schedule prediction and had greater time complexity, however. SAM performed more poorly when applied to long-duration studies, it failed to generalize when predicting performance across timescales that varied from those used to fit the model, and it had the greatest time complexity.

5. General discussion

The spacing effect is one of the most widely obtained results in psychology research. In addition to revealing fundamental details about memory, research on the spacing effect has substantial implications for education and training. Indeed, several recent reports identify distributed practice as one of the most effective learning techniques (Dunlosky et al., 2013; Pashler et al., 2007; Roediger & Pyc, 2012). Rigorous theory is needed to bridge the gap between laboratory experiments and educational enterprise. Computational models of knowledge acquisition and retention provide such theory.

To harness the potential of these models, it is necessary to take stock of their strengths and limitations. To that end, we proposed a set of theoretic criteria for evaluating models of the spacing effect. These began with the basic effects of massed versus spaced practice on knowledge acquisition and retention, and they went on to include how myriad factors such as length of retention interval, amount of practice, and number of re-learning sessions modulate the basic effect. The resulting collection of “spacing effects” is complex, yet as our simulations showed, they can be accommodated by computational models that
posit relatively simple mechanisms. Predictive Performance Equation captures the spacing effect using spacing-dependent decay and P&A captures the spacing effect using activation-dependent decay. SAM, a hybrid model, captures the spacing effect using deficient processing and contextual fluctuation. SAM includes a third mechanism, retrieval-dependent trace updates, to produce the non-monotonic effect of spacing on retention.

In addition to theoretic criteria, we proposed a set of criteria for evaluating the applied potential of models of the spacing effect. Most of the applied criteria (i.e., 1–4) deal with issues of scalability and generality. The basic message is that to be of practical use, a model of the spacing effect must go beyond narrowly accounting for an experiment finding: theoretic adequacy is necessary but not sufficient. Some of the applied criteria (i.e., 3 and 5) depend on the intersection between model properties, statistical methods, and hardware. For example, successful out-of-sample prediction (Applied Criterion 3) partially depends on adopting fitting methods that adequately penalize and constrain overly flexible models. Yet simpler models are less prone to overfitting. Model run time (Applied Criterion 5) is platform-dependent. Yet a model’s time complexity will ultimately limit the range of inputs that it can feasibly handle. None of the applied criteria are specific to computational models of the spacing effect—all are relevant to determine the application potential of computational models of other cognitive phenomena.

One limitation of SAM is that parameters that control the rate of contextual fluctuation bound the timescale over which the model operates. This potentially limits the model’s ability to account for performance across widely varying timescales characteristic of real-world learning (Applied Criterion 2). The multi-scale context model (MCM; Mozer et al., 2009) instantiates a contextual fluctuation mechanism as in SAM, but it does so using pools of units that decay at different rates. Increasing spacing shifts the representation to units with a lower decay rate, thereby scaling contextual fluctuation to the appropriate timescale. This leaves open the possibility that MCM, or some variant thereof, could extend a core mechanism in SAM, contextual fluctuation, to learning across multiple timescales. Another attractive feature of MCM is that it is highly predictive; one published study showed that the model can be calibrated with data from an acquisition session and used to predict retention after varying ISIs and RIs (Mozer et al., 2009; Applied Criterion 3). Furthermore, a model based on MCM improved learning outcomes in a semester-long in-class study (Lindsey et al., 2014; Applied Criterion 2). Multi-scale context model should be explored further, and its theoretic scope and application potential rigorously evaluated.

5.1. Comparison between PPE and P&A

The similar performance of PPE and P&A across the five theoretic criteria and the 14 datasets raises the question, how do the models relate to one another? The predecessor to PPE, the General Performance Equation, was derived from the base-level activation equation in ACT-R (Anderson & Schunn, 2000). P&A also builds on the base-level activation equation in ACT-R. As a consequence, both models account for the first-order effects of amount of practice and elapsed time since practice in similar ways. Additionally, spacing
impacts memory performance in P&A and PPE in a similar way—through its effect on decay.

An equally important question raised by these comparisons is what makes the models distinct. Predictive Performance Equation and P&A differ in at least two meaningful ways, which may allow for future experiments to distinguish between them. First, decay in PPE is spacing dependent, whereas decay in P&A is activation dependent. Other factors besides spacing in ACT-R affect an item’s activation, and would therefore be expected to accelerate acquisition and reduce retention. For example, retrieval cues may spread activation to a chunk, making it accessible from memory. As a side effect, subsequent repetitions will be stored with higher decay rates, impairing retention. If an experiment were conducted that varied spacing and number of retrieval cues presented during initial learning, P&A predicts that both factors would affect retention, whereas PPE predicts that only spacing would affect retention.

Second, PPE and P&A make different predictions regarding the impact of initial spacing on subsequent re-learning (Walsh et al., in press). Over long intervals between study and test, a large value of elapsed time (T in Eq. 2) may 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 the brief elapsed time since repetitions from within the session. The benefits of practice and decay (N and d in Eq. 2) will re-emerge 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 re-emerge once knowledge is reactivated by re-learning. P&A predicts that spacing will not affect—or may even limit—gains from re-learning trials. If an item initially studied in a spaced manner has greater activation at the start of a retention test, subsequent repetitions of that item will be stored with higher decay rates. In other words, although spacing may improve retention in P&A, it will limit gains from re-learning trials. If an experiment were conducted that measured the rate of re-learning of forgotten items after a long enough retention interval, P&A predicts that all items will be relearned at the same rate, whereas PPE predicts that items initially studied in a spaced manner will be relearned more quickly.

5.2. Other criteria

Although the theoretic criteria we identified span a diverse set of experiment materials and manipulations, they are still relatively narrow. Ultimately, spacing effects are only one type of memory-related phenomenon. Other memory-related phenomena have equally strong implications for education and training. For instance, tests are usually seen as a means for measuring mastery. However, the act of taking a test produces large learning gains, even relative to other forms of study (Roediger & Karpicke, 2006). Additionally, self-explanation, or reasoning about why something is, improves retention (Berry, 1983; Chi, Bassok, Lewis, Reimann, & Glaser, 1989). These and other memory-related phenomena, which may interact with the spacing effect (Rawson, Vaughn, & Carpenter, 2015),
fall beyond the scope of PPE and other spacing effect models, but are equally vital to education and training.

Most of the experiments in Table 4 involve a single spacing-related phenomenon drawn from the five theoretic criteria. Uncovering spacing-related effects in such a piece-meal fashion is a necessary first step. However, from a theoretic perspective, demonstrating multiple effects with one group of participants and in one experiment provides a stronger test of a computational theory. A model may capture effects observed in different experiments through systematic variations in parameter estimates. Using one group of individuals and one task—and hence one set of parameter estimates—provides a clearer picture of whether the model’s mechanisms adequately capture the multitude of effects.16

The applied criteria we identified may be too narrow as well. One thing missing from the list is ease of understanding and implementing the model. To be of use to practitioners, computational models of learning and forgetting must be more accessible. This involves sharing source code along with model descriptions and creating software to automate model calibration and prescription (i.e., Jastrzembski, Rodgers, & Gluck, 2009; Lindsey et al., 2014). Another thing missing is quantification of uncertainty. Practitioners must be able to calculate model predictions as well as confidence intervals surrounding those predictions to properly mitigate risk when making education and training decisions.

We propose the criteria identified in this article as a starting point for defining the scope of computational models of learning and retention, and as a progress check for those working at the intersection of cognitive modeling and the learning sciences. Additionally, we propose them to place equal emphasis on theoretic adequacy, which is typically the focus of empirical research, and application potential, which is often neglected. Only by addressing both can we translate computational models of learning and memory to educational practice.

Notes

1. This represents the proportion of elements encoded during the first presentation ($x$) plus the proportion of remaining elements encoded during the second presentation ($x(1 - x)$).

2. Retention depends on the length of the ISI and the RI (Cepeda et al., 2008). Benjamin and Tullis (2010) did not evaluate superadditivity jointly on the basis of ISI and RI because there were too few empirical studies to do so. We use computational simulations to address this issue in a later section.

3. Admittedly, our theoretic evaluation in Section 4 is based on goodness of fit.

4. Spaced practice sometimes speeds learning in motor skill acquisition. This may occur because spaced practice reduces fatigue during motor skill performance. Other times, spaced practice slows initial learning and impairs subsequent retention (Paik & Ritter, 2015). This may occur when massed practice is needed to convert declarative knowledge to a procedural form during initial learning.
5. A low value of $c$ produces substantial gains from initial practice and diminishing gains from further practice. Model simulations across the 14 data sets reported in the next section produced comparable results for many values of $c$; for values ranging from 0.01 to 0.20, the average root-mean-squared error (RMSE) across the data sets changed by $<2\%$. Given the equivalently good results across this range, we fixed this parameter at the median value.

6. Model simulations across the 14 simulated data sets reported in the next section yielded comparable results for many values of $x$; for values ranging from 0.3 to 0.9, the average RMSE across the data sets changed by $<5\%$.

7. Lag prior to the first presentation of an item ($\text{lag}_1$) is $\infty$, and the decay rate following the first presentation of an item (i.e., $j = 1$) is simply $b$ (Eq. 5). Lag following the second presentation (i.e., $j = 2$) equals the difference between the times of the first and second presentations ($t_2 - t_1$).

8. Decay rates ($d_n$ in Eq. 5) were greatest for the massed schedule, followed by the intermediate and spaced schedules (decay at $R1 = -0.070, -0.062, -$0.056; decay at $R2 = -0.069, -0.061, -0.056$). These variations in decay rate accounted for about 80% of the variance in retention performance at $R1$ and $R2$.

9. The average decay rates of the 16 instances of CVC pairs at the end of the acquisition phase were greatest for the massed schedule ($-0.71$), followed by the intermediate schedule ($-0.63$) and the spaced schedule ($-0.59$).

10. Based on the simulations, we calculated when items left the STS and how many times they were re-encoded in each condition. On average, items were re-encoded the fewest number of times in the massed condition (7.8), followed by the intermediate condition (10.9) and the spaced condition (12.2).

11. By allowing SAM’s contextual fluctuation parameters to differ for each RI, the model produced a range of monotonic and non-monotonic retention functions with peaks at different ISIs. However, there is no principled basis to vary these parameters by condition.

12. Retention was lower when the RI was 7 days versus 1 day, but RI did not interact with the other factors. Following Pavlik and Anderson (2005), we averaged observations and model results over the two RIs.

13. Pavlik and Anderson’s study included multiple trials with feedback per item during the test session. Model results are shown only for the first exposure to each item during the test session, but parameter estimates are based on data from all trials.

14. P&A and SAM also predicts a subadditive effect of repetition at short ISIs.

15. Experiment 1 consisted of four sessions (Days 1, 3, 8, and 10), Experiment 2 consisted of five sessions (Days 1, 3, 8, 10, and 38), and Experiment 3 consisted of four sessions (Days 1, 3, 8, and 10).

16. This is somewhat in tension with the ideal test of a model’s application potential, which would involve having a group of individuals perform multiple different tasks to evaluate the model’s suitability for each task (i.e., Applied Criteria 4).
References


Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Model parameter estimation.