Abstract

Memory should make more available things that are more likely to be needed. Across multiple environmental domains it has been shown that a system that behaved in this way would match memory effects evolving repetition, delay, and spacing (Schooler & Anderson, 2016). We examined the patterns with which words appear in two large-scale data sets: tweets from popular sources and comments on popular subreddits. These sources are sufficiently large that they enable accurate measurement of how detailed patterns of past appearance predict probability of occurring again. None of a candidate set of mathematical models of memory does very well at predicting the observed patterns. Anderson & Milson (1989) had proposed a model of the environment which assumes that there is a distribution of desirability of items, that their desirability decays according at different rates, and that items occasionally have revivals to their original desirability. While Monte Carlo simulation of the original model, which assumed exponential decay, does not fit these two sources particularly well, simulations of a model which assumes power-law decay does. A mathematical model of human memory based on the revised Anderson and Milson environmental model did better than other models at predicting the environmental data and a wide range of behavioral studies focused on how spacing of items affects probability of recall. This model also predicts latency effects associated with presentation patterns at frequencies that are too high to study with probability of recall.

Keywords:
Memory, Rational Analysis, Environmental Statistics, Spacing Effect
1. Introduction

Rational analysis (Anderson, 1990) has been a productive approach for the past three decades in cognitive science, addressing a variety of topics in different, novel ways (e.g. Chater & Oaksford, 1999; Gershman et al., 2015; Griffiths et al., 2015; Kemp & Regier, 2012; Lewis et al., 2014). With similarities to Marr’s computational level (Marr, 1982), it proposes that an abstract specification of human behavior can be derived as an optimal solution for achieving human goals in an uncertain environment and within computational limitations. In the early analyses of Anderson (1990), the emphasis was on cognition as emerging as a response to the statistical structure of the environment with minimal commitments about cognitive mechanisms. Many of the more recent approaches (see Lieder & Griffiths, 2020), have focused on how computational limitations shape cognition in the spirit of ecological rationality (Todd & Gigerenzer, 2012).

Simon (1989) made an early and cutting criticism of rational analysis, particularly with its emphasis on the environment. Pointing to issues with the use of rationality in economics, he argued that rational analysis lacked any real predictive power. Any behavior is optimal in some imaginal world and he argued that all rational analysis was doing was imagining a world in which human behavior would be optimal. In his thesis, Schooler (1993) produced what is still the best response to Simons critique. Anderson & Milson (1989), henceforth A&M, had claimed that memory was shaped to make most available what was most likely to be needed. In analyses of three environments that make memory demands on humans (word use in New York Times headlines, word use in caregivers speech to children, and sources of email messages) Schooler showed that a system that made most available what was most likely matched human memory.

Figure 1 illustrates some of the effects as displayed in Anderson & Schooler (1991) looking at what can be predicted from the pattern of appearance of a word over the last 100 days in New York Times headlines – similar patterns were found in caregiver speech to children and email messages. Human forgetting has been characterized (e.g. Wixted & Ebbesen, 1991) as a power function – probability of recalling a memory decreases as a power function of the amount of time that has passed. Figures 1a and 1b illustrate that this same relationship holds in the environment, plotting probability of a word appearing in a New York Times headline in the next day as a function of the number of days since it last appeared. Part (a) shows a function that initially drops rapidly but also rapidly decelerates in its descent. Part (b) shows that this function becomes linear when both
Figure 1: Anderson & Schooler (1991)'s predictive patterns in the New York Times: (a) Probability of a word occurring on the next day as function of how long ago it last occurred; (b) Log-log transform of part of part a; (c) Joint effect of frequency of occurrence in the last 100 days and delay since last occurrence; (d) Items that occurred twice in the last hundred days effect of lag between the two occurrence and time since last occurrence.
axes are log transformed\textsuperscript{1} the signature of a power function.

Figure 1c shows how the frequency with which a word has appeared in the last 100 days combines with delay since the last appearance of the word. The three frequency bands are approximately parallel linear functions of delay on a log-log plot. Moreover, the curves are approximately spaced according to the log of the frequency, giving rise to the observation that odds is approximately a product of a power function of practice (N) and a power function of delay (T):

\[ \text{Odds} = A \times N^c \times T^{-d} \]  

(1)

Applying this equation to learning, \cite{Anderson&Schunn2000} called this function the General Performance Equation (henceforth, GPE).

Figure 1d is for words that appeared just twice in the last 100 days and shows the interaction between the spacing of the two occurrences and the delay since the last occurrence. The different curves reflect different delays and are ordered with the highest probability for the shortest delay. Reflecting the classic spacing effect \cite{Cepedaetal2006} in human memory, the drop in probability with delay is greatest for items that occurred twice close together. As we will see, it proves challenging to understand how the spacing effect in Figure 1d relates to the apparently independent effects of practice and delay in Figure 1c.

Since this paper will be doing a similar analyses, it is important to understand how the data in displays like Figure 1 are calculated. In all cases the analysis is examining how the pattern of occurrence of a word like Qaddafi in the last 100 days of New York Times headlines predicts the probability it occurred on the 101st day. This analysis comes from 730 days of headlines that that span 1987 and 1988. To perform this analysis, a 101-day window is slid day-by-day across the period and results are averaged. Some points like the probability of occurring on the next day (1 on the x-axis in Figure 1a), are averages of all words meeting that exact pattern (in this case, all words that appeared on the 100th day for all of these windows). In other cases, to get reliable data patterns are averaged together. This is true of the later points in part a and b of Figure 1, many of the points in part c, and all of the points in part d. One might want to identify accurately the probability of re-appearing for all $2^{100}$ patterns of past occurrence, but this is obviously a desire that cannot be satisfied with any plausible amount of data.

\textsuperscript{1}Log odds is chosen so that the value is not bounded above by 1. At low probabilities there is little difference between probability and odds.
This paper will provide more information on how past patterns of occurrence predict probability of appearing now. The size of their samples did not allow Anderson & Schooler (1991) to perform a thorough examination of the interaction between recency, frequency, and spacing. Also, they only traced out temporal effects going back 100 days or utterances which limits an accurate characterization of the effects of time (the effects of time prove not to be a simple power function). This paper will examine much larger modern data sets to measure in more detail what the patterns are in the environment and examine patterns over longer time windows.

This paper’s more detailed analysis of environmental patterns will have three major dividends. First, it will provide a stronger test of the fundamental claim of rational analysis that memory is adapted to the statistical structure of the environment. Second, it will enable explorations of human memory in ways that are not possible behaviorally. Third, capitalizing on the second, we will test existing theories of memory and develop a better theory.

2. Studies of Environmental Patterns

Below we describe two modern data sets that are much larger than the data sets used by Anderson & Schooler (1991). They have distinct properties that allow each to address some interesting questions about human memory. Like Schooler’s data from the New York Times and caregiver speech, they involve a sequence of texts. We will be examining the statistical patterns of the strings in these texts (like Qaddafi in the New York Times headlines) and how these patterns predict the appearance of the strings in future texts. One sequence of texts will be tweets that appear across years from highly followed individuals. The other sequence of texts will be comments that appear in popular subreddits on a single day. In both cases the question of interest is how the appearances of a string in prior texts in the sequence predicts its probability of occurring in the next text. The underlying assumption is that each appearance of a string is a demand on the memory of the follower to understand the referent of the string. Despite their differences these data sets will prove to have some strong commonalities that allow us to test and refine theories of memory, particularly focused on the spacing effect.

2.1. Twitter Database

The first data is a subset of the data in Stanley’s dissertation (Stanley 2014; Stanley & Byrne 2016) on predicting hashtag usage. The subset involved the tweets of the top 500 English tweeters measured by number of followers as of the collection of the data (Jan
It contained all their tweets from near the beginning of Twitter (July 11, 2007) to the collection date (Jan 7, 2014). Each tweet from a source posed memory demands on someone following the tweeter. For instance, here is a tweet from barackobama:

This debate is not just about numbers. It’s a set of major decisions that are going to affect millions of families.

A reader of this tweet would have to know what debate refers to, but every content word is a demand on memory to retrieve its meaning. We stripped high frequency functor words from the tweets, limited all tweets to a vocabulary of the top 20,000 words, and eliminated repeated words in the string. Hashtags survived if they had high enough frequency. The surviving strings from the example tweet become:

'debate' 'just' 'numbers' 'set' 'major' 'decisions' 'going' affect 'millions' 'families'

This process yielded 1,038,632 tweets averaging 7.5 unique strings.

We examined how the pattern of appearance over 1000 tweets predicted the probability of occurring in the 1001st tweet. Just 361 of the original sources had in excess of 1000 tweets and could be used. They averaged 2722 tweets. Thus, there was an average of 1721 1000-tweet windows for these tweeters that could be used to predict strings in the next tweet. Each word that appears at least once in a 1000-tweet set will define a pattern of appearance to predict whether it occurs in the 1001st tweet. There are over slightly less than 1.2 billion such patterns.

Even though there are more than a billion patterns, there are many more (\(2^{1000}\)) possible patterns than observations. Still the size of the data set allows for examining detailed information about how past patterns predict future appearance of a string. Figure 2 displays analyses of these that highlight the effects of recency, frequency, and spacing. Part (a) shows how probability of a string appearing in the 1001st tweet increased with the number of tweets it had appeared in and decreased with the number of tweets since it last appeared. The nth frequency band includes 2n-1 specific frequencies. We similarly aggregated delays into bands that had 2n-1 observations and plotted it on the x-axis as the average delay in that band. This means, for instance, that the fourth point on the third curve aggregates all cases with 5-9 appearances of the string in the last 1000 tweets and the most recent appearance between 10-16 tweets ago. There are 4,134,835 instances

\(^{2}14.5\text{ percent of the strings in the 1001st tweet have not appeared in the previous 1000 tweets.}\)
Figure 2: Predictive patterns in the Twitter data: (a) Joint effect of frequency of occurrence in the last 1000 tweets and delay since last occurrence; (b) Items that occurred once or twice in the last 1000 tweets – effect of lag for twice occurring.

contributing to this point and in 35,117 of these cases that string occurred in the next (1001st) tweet for a percentage of 0.85%. That value is plotted on the 4th curve at x=13 which is the mean of 10-16. To better reveal what is happening, both the x and y axes are plotted in log units. Cases are only plotted that have at least 5000 observations. The first four frequency bins have all observations plotted out to the longest delay bin, but the larger frequency bins are missing some of the longer delays. The highest frequency band (197-225) only has observations out to the fifth lag bin (17-25) where there are 11,296 cases.

The patterns correspond somewhat what Anderson & Schooler (1991) reported – parallel linear functions of delay. The best-fitting function of the form

$$\log(\text{probability}) = a + b \log(\text{Frequency}) + c \log(\text{Lag}).$$

has the values $a = -3.67$, $b = .54$, and $c = -.65$ implying the following GPE (see Equation 1 – ignoring for the moment the distinction between odds and probability)

$$\text{probability} = .026 \times \text{Frequency}^{.54} \times \text{Lag}^{-65}.$$
to Figure 1c, which covers this range. However, this pattern does not fully generalize to greater frequencies and longer delays.

Part b of Figure 2 is focused on those cases where an item has just occurred twice and investigates the effect of the lag between those two occurrences, breaking that lag into bands. This replicates the pattern found by Anderson & Schooler (1991) but in greater detail. At a short delay, a short lag is better but at long delays the greater lag is better. For all of the lag bands, the effect of delay is negatively accelerated, which consistent with the low frequencies in Part (a). Figure 2b also shows the N=1 curve from part (a). At long lags there is no difference between a single presentation (N=1) and two back-to-back presentations (Lag=1). It is as if those two presentations have been merged into one.

The x-axis in Figure 2 is number of intervening tweets. These tweets are tagged for when they were sent, allowing us to address the perennial question in memory research: whether the passage of time promotes forgetting or whether forgetting is really caused by intervening events that interfere with the original memory (Wixted, 2004). The time since the last tweet containing a string, if that string occurred in the last 1000 tweets, varies from 0 to 1860 days with a median of 36.1 days – 5.5\% of the cases are less than a day, 25.4\% less than 10 days, 72.0\% less than 100 days, and 99.9\% less than 1000 days. Not surprisingly, number of intervening tweets and number of days are correlated (r=.497), but the correlation is not so strong that they cannot be separated. Also not surprisingly, both are weakly correlated with frequency (order r=-.054, time r=-.027), which has a strong effect of its own. We investigated whether time or lag was predictive of the string occurring in the next tweet for different frequency bands. Figures 3a and b look at cases with a single mention of the string in the last 1000 tweets, with the x-axis either displaying the effect of intervening days for different bands of number of intervening tweets (part a) or the effect of intervening tweets for different bands of number of intervening days (part b). The strong effect is number of intervening tweets. Figures 3c and d look at cases of intermediate frequencies 2-9 and support the same conclusion. The same pattern appears in Figures 3e and f for frequencies greater than 9. While these results do not support the conclusion that passage of time has no effect, it does justify a focus on number of intervening texts.

---

3However, the frequency bands refer to number of occurrences in the last 1000 tweets, not the last 100 days as in Figure 1c.
Figure 3: Contribution of intervening time versus number of intervening tweets plotted for different frequency ranges.
2.2. Reddit Database

The Reddit data provides a test of whether the same patterns will emerge in a sequence of texts that is produced in quite different way. Most of the analyses will be of the posts on the top 501 subreddits (by subscribers) on two days, April 23 and May 5, 2021. For each subreddit we took the perspective of someone browsing that subreddit and processing the strings in comments like:

I received 100% attendance in 4th grade and got a free ice cream dessert at Ponderosa.

We assumed that the subscriber would read the first 25 discussions listed on the subreddit. Reddit has an algorithm that orders these discussions both by recency and votes of the users. While these discussions can go on very long, by default Reddit makes available no more than 200 comments although one can always click to see more (we assume our subscriber does not click). Comments are hierarchically organized with the top-level being responses to the original story and below these are comments on the comments and below that comments on those, etc. By default the comment depth is 10 although again one can click to see more. The order of the comments at a level is determined by voting of the users (apparently no effect of recency). The net effect of these Reddit algorithms is that the order in which our hypothetical subscriber reads things is weakly affected by recency of posting and strongly affected by group popularity. This contrasts with the order that tweets are read by the follower of a tweeter, which is determined by the order the tweeter sends them.

The maximum number of comments that our hypothetical subscriber would read would be 25 stories times 200 comments equals 5000. In fact, the number of comments on a subreddit ranged from 104 to 4691 with a mean of 1,131. 439 subreddits had a least 1001 comments with a mean length of 1832. Again, we stripped out any functor words or any strings that were not among the 20,000 most used strings. We also limited any comment to the first 100 strings. In total, there were 1,133,182 comments averaging 14.85 strings in length. The number of 1000-string patterns is similar to Twitter – slightly more than 1.2 billion patterns.

Despite their common size, the two sources are quite different in how they were created. The Twitter data set describes the followers of a tweeter over years. The Reddit data set reflects the experience of a user in one sitting. The tweets are ordered by when they were sent while the ordering of the comments is determined by the Reddit algorithms strongly reflecting group popularity. The topic usually changes at least a little with each
Figure 4: Predictive patterns in the Reddit data: (a) Joint effect of frequency of occurrence in the last 1000 comments and delay since last occurrence; (b) Items that occurred once or twice in the last 1000 comments – effect of lag for twice occurring.

discussion in Reddit, while any topic change in Twitter is at the tweeter’s discretion. Both the Twitter and Reddit data sets are artifacts shaped by humans but they are quite different in their shaping. To the degree they reflect similar statistics, this is evidence for the similarity of the demands faced by human memory in trying to make most accessible what is most needed.

Figure 4a displays the frequency and recency effects to compare with Figure 2a. Figure 4b displays the spacing effects to compare with Figure 2b. The patterns are strikingly similar to Figure 2. Parts (a) of Figures 2 and 4 have mean percentages of 2.80% and 2.85% with standard deviations of 4.77% and 4.64%. In probability scale the numbers in Figure 2a and 4a are correlated .997 and in log probability they are correlated .986. Parts (b) of Figures 2 and 4 have mean percentages of 0.33% and 0.39% with standard deviations of 1.10% and 1.12%. In probability scale they are correlated .973 and in log probability they are correlated .990. The similarity in magnitudes is in part due to the decision to seek out more than a billion patterns of length 1000 involving 20,000 strings. The similarity in the patterns, on the other hand, speaks to the ubiquitous way experiences are shaped. Joined with the original Anderson & Schooler (1991) analyses and other subsequent analyses (see Schooler & Anderson (2016) for a review) the ubiquity seems profound.

The Reddit data allows a check of a couple of patterns that have been investigated in studies of spacing. A number of studies of memory starting with Bahrick (1979) have examined effect of spacing over many days (these will be discussed more later in the paper). All include a condition where most or all of the study is massed on one day and contrast this with study split over days. In all cases they look at retention at a number of days from the last study. To investigate the needed patterns, we collected Reddit
comments on a third day, July 17th. We compared cases of having N occurrences of a string on April 23rd and N on May 5th with cases having all 2N concentrated on May 5th. In terms of the likelihood on the string occurring on July 17th, having all occurrences on May 5th would have the recency advantage, but having occurrences spread over the two days has the spacing advantage. The memory studies generally find spacing to dominate. These studies typically have some nominal spacing between repeated practices on the same day. Therefore, we divided each of the days into 3 equal parts and asked how having N occurrences in the first and third parts of May 5 and 0 occurrences elsewhere, compared to having N cases either in the first part of April 23 and the third part of May 5 or the third part of April 23 and the first part of May 5 (again 0 occurrences elsewhere). We assessed these effects separately on the first and third parts of July 17. There were enough observations to answer the question for N = 1, 2, and 3: In each case, 2N on May 5 was non-significantly better than N on April 23 and N on May 5 in terms of percentage of occurrences on July 17 (27.5% vs 26.7% when N= 1 for 179 subreddits – t(178) = 0.75; 44.4% vs 42.2% when N= 2 for 170 subreddits – t(169) = 1.15; 51.7% vs 52.8% when N= 3 for 76 subreddits – t(65) = 0.22). Thus, these comparisons show no sign of the behavioral result of spaced practice across days being better.

On closer inspection one can question whether these are the right contrasts. Each study in Bahrick’s experiment involved training the item to a criterion in a drop-out procedure. Not surprisingly, the number of presentations for the first study is greater than for later studies and the number of presentations for a second study on the same day is less than for a second study on a later day. A number of studies subsequent to Bahrick have tended to control the number of exposures after the first drop-out training but still one can wonder how much subjects attend to further exposures on the same day after reaching mastery. To approximate these considerations in the Reddit data, we contrasted cases that had 3 occurrences in the first part of May 5 and 1 in the third part of May 5 (massed case) with 3 occurrences in the first part 1 of April 23 and 2 in the second part of May 5 (spaced case). This approximates the situation where the first drop-out has the most exposure, further trials on the same day the least, and further trials on a later date somewhere in between. This contrast yields an effect corresponding to the behavioral result 50.2% for spaced and 38.3% for massed averaged over 164 subreddits – t(163) = 4.75.

The Reddit data also affords a check of another question about effects of distribution over multiple days that has been studied: How does the pattern on study of an item during one day predict its retention on a later days and how does information drop off with the
Figure 5: Effect of frequency and Range in the first 1000 comments of April 23 on probability of occurring in the last 1000 comments of (a) April 23, (b) May 5, and (c) July 17. (d) Effect of frequency and Range in Twitter.
passage of time? Restricting ourselves to 74 subreddits that had at least 2000 comments on April 23rd, Figure 5 displays how the pattern of appearance in the first 1000 comments on that day predicted probability of occurring in the last 1000 comments of that day, the last 1000 comments on those subreddits on May 5, and the last 1000 comments on those subreddits on July 17th. The x-axis in parts (a) – (c) reflects the number of occurrences of a string in the first 1000 comments on April 23. The different points and lines reflect Range, which is the number of comments that occurred from the first to last mention of an item, including the first and last mentions. An item that has a Range of 1 necessarily has frequency of 1 and so it is a single point in the figures. Similarly, an item which has a Range of 2 must have a frequency of 2 (i.e. the two appearances of the string were in adjacent comments) and so it also a single point. Holding frequency constant, Range has a large effect on probability of occurring on that day and on subsequent days. The figure also shows that frequency has a quite small effect for items that have a small Range. It is as if all those presentations in that small Range have been collapsed into a single data point.

Part (d) of the figure presents the corresponding analysis for Twitter restricting ourselves to the 306 tweeters that have more than 2000 tweets. We asked how a string’s pattern of appearance in the first 1000 tweets predicted how likely the string was to occur in the last 1000. This shows a pattern very similar to the Reddit patterns. Overall, Range has a larger effect than frequency and it will prove the critical piece of the model described later in the paper.

3. Fitting Environmental Data

The rational hypothesis would imply that it should be able to map a correct model of memory recall onto the environmental data. Table 1 reports fits of various models of memory to the data from these two environmental sources, starting with the fit of the GPE (applied to odds as Equation 1), which sets an aspiration level for models with stronger psychological claims. Figure 6 compares the average of the Twitter and Reddit data (top) with the GPE predictions (bottom). Data and predictions are for all cases in Figure 6 with more than 5000 observations (326 data points in Figure 6a and 187 in Figure 6b). In fitting this and all other models we minimized the sum of squared deviations from

---

4Given the evidence in the Twitter data that the critical factor is number of intervening texts and not time, we will interpret what is referred to as time in these models as number of intervening texts and not clock time.
Figure 6: Predictive patterns in the Combined data: (a) Joint effect of frequency of occurrence in the last 1000 texts and delay since last occurrence; (b) Items that occurred once or twice in the last 1000 texts effect of lag for twice occurring; (c) and (d) predictions of the best GPE model for parts (a) and (b).
Table 1: Fit of Various Models to the Environmental Data

<table>
<thead>
<tr>
<th>Measures of Fit</th>
<th>RMSE</th>
<th>r^2</th>
<th>Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Mathematical Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPE</td>
<td>0.583</td>
<td>0.888</td>
<td>c=.582  d=.617 A=.021</td>
</tr>
<tr>
<td>ACT-R</td>
<td>0.825</td>
<td>0.784</td>
<td>d=.798  b=.040</td>
</tr>
<tr>
<td>P&amp;A</td>
<td>0.684</td>
<td>0.847</td>
<td>c=.453  a=.758 b=.053</td>
</tr>
<tr>
<td>PPE</td>
<td>0.549</td>
<td>0.898</td>
<td>x=.8699 c=.619 b=.536 m=.186 A=.018</td>
</tr>
<tr>
<td>MCM</td>
<td>0.578</td>
<td>0.888</td>
<td>µ=.032  υ=1.111 ω=.704 ξ=.978 A=.029</td>
</tr>
<tr>
<td>AMPE</td>
<td>0.398</td>
<td>0.947</td>
<td>a=214   b=1401 tP=15.2 gP=1565</td>
</tr>
</tbody>
</table>

(b) A&M Simulations of the Environment with Different Decay Functions

<table>
<thead>
<tr>
<th>Parameters:</th>
<th>Exponential</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>v, b: parameters defining distribution of desirabilities</td>
<td>v=.164  b=.139 α=0.035  β=333  A=.704</td>
<td></td>
</tr>
<tr>
<td>a: scale parameter for desirability</td>
<td>a=.162  b=.482 α=.002  β=600  A=.751</td>
<td></td>
</tr>
</tbody>
</table>

Parameters:

GPE:  c: Exponent for number of presentations
      d: Decay exponent for time since last occurrence
      A: Scaling parameter for odds

ACT-R: d: Decay exponent for each presentation
       b: Scaling parameter for odds

P&A:  c: decay parameter multiplied by activation
       a: minimum decay
       b: Scaling parameter for odds

PPE:  x: power in estimation of decay time
       m: decay parameter multiplied by lag estimate
       b: minimum decay
       A: Scaling parameter for odds

MCM:  µ: scale parameter for decay rate
       ν: decay rate parameter raised to power
       α: scale parameter for weight strength
       ξ: weight parameter raised to power
       A: Scaling parameter for odds

AMPE a: scale parameter for desirability
       b: scale parameter for decay
       tP: prior time in estimation of decay time
       gP: prior gap in estimation of effective interval

A&M  v, b: parameters defining distribution of desirabilities
     α: parameter defining distribution of decays
     β: parameter controlling rate of revival
     A: Scaling parameter for odds
log probabilities. The choice to focus on log probabilities is to emphasize the lower probabilities. Any memory system would make available highly probable items. The challenge is how to treat the less probable. The GPE model captures the major effects of number of occurrences and number of texts (tweets or comments) since last occurrence – these are the factors N and T the GPE equation. However, it does not capture other features of the data, most apparently spacing effects. Despite its glaring deficiencies, it proves to outperform a number of memory models, including ones that address spacing effects.

3.1. ACT-R

Most models of human memory make no strong claims about a relationship to statistics in the environment, but this is not true of the ACT-R theory (Anderson & Lebiere, 1998). Activation of declarative memories in ACT-R is supposed to reflect log odds of the memories being needed. One component of activation, called base-level activation, is supposed to reflect the effect of the past history of a memory on its log odds. Anderson & Schooler (1991) suggested that the following equation for predicting the logs odds of an item appearing given its past history:

\[
\text{Odds} = b \times \sum_{j=1}^{n} t_j^{-d}
\]

(2)

where the summation is over the n times the memory appeared in the past, \(t_j\) is how long ago the jth appearance was, \(d\) is parameter reflect rate of decay with time and \(b\) is a scaling parameter. This equation was directly incorporated into ACT-R as the equation for base-level activation of a memory i:

\[
B_i = \log(\sum_{j=1}^{n} t_j^{-d}) + B.
\]

The ACT-R model only requires two parameters, \(d\) and \(b\) in Equation 2, to fit the data from the environment. Figures 7a and 7b show the predictions of this model. The

---

5This typically required converting an odds prediction of a model like GPE into log probability. There is very little difference between odds and probability for low probabilities.

6Petrov (2006) showed that the summation in this equation could be efficiently approximated by representing the average of the j most recent presentations and then adding in an average effect of the remaining:

\[
\sum_{i=1}^{n} t_i^{-d} = \sum_{i=1}^{j} t_i^{-d} + \frac{n-j}{1-d} \times \frac{t_n^{-d} - t_j^{-d}}{t_{n-i}^{-d}}
\]

We used this equation with \(j = 2\) so that we could precisely predict the effects of lag when there were two presentations.
Figure 7: (a) and (b) predictions of the best-fitting default ACT-R model. (c) and (d) predictions of the best-fitting P&A model.

ACT-R model captures major trends of frequency and recency, but overall it fit is not as good as the GPE (Table 1—it does have one less parameter). The delay functions in Figure 7a are basically straight lines on this log-log scale and so fail to capture the empirical shape of the curves, especially the precipitous drop of the high frequency curves at long lags reflecting the effecting of close spacing. Figure 7b confirms the failure to capture spacing effects. The lag between the two occurrences does show an effect at short delays in Figure 7b because of the greater delay to the first occurrence in the pair, but the curves for different lags just converge rather than cross over at long delays.

3.2. P&A

Pavlik & Anderson (2005), henceforth P&A, introduced an elaboration to the ACT-R model that captured many spacing effects on recall probability. As in ACT-R, activation of a memory was determined by a sum of decaying components, but P&A, proposed different decay rates for different presentations. Cast as a prediction about odds (like Equation 2) their model would be:

\[ \text{Odds} = b \times \sum_{j=1}^{n} t_j^{-d_j} \]  

(3)
The decay for the first presentation was fixed:

\[ d_1 = a \]  

(3a)

But later presentations would have greater decays to the extent that the item was active at the point of presentation:

\[ d_n = c \times \sum_{j=1}^{n-1} t_j^{-d_j} \]  

(3b)

Parts (c) and (d) at the bottom of Figure 7 show the fit of this model, which is better in terms of measures of fit than the standard ACT-R model (Table 1). It definitely captures the precipitous drop of high frequency presentations at long lags (part c) and shows some spacing effect for twice presented items (part d). However, it also does not do as well as the GPE model, which has no spacing effect. Perhaps most dramatically, it underestimates the benefit of many presentations at short delays. This is because the occurrences are rapidly decaying and the dominant components are just the first presentation (slow decay) and last presentation (short delay).

3.3. PPE

Given that neither the GPE nor either version of ACT-R in Figure 7 capture the environmental patterns, we considered two other models that were specifically designed to address spacing in recall probability. These models make no claims about the environment, but to apply them we will assume simple monotonic relationships between probability of recall and probability of occurring in the environment. Probability of remembering something should be higher than probability of occurring in the environment. For instance, an item that has a 10% chance of occurring in the next text should probably have close to 100% probability of being remembered in an adaptive memory.

Walsh et al. (2018) proposed what they call the Predictive Performance Equation (henceforth PPE) as a refinement of the GPE. Their model is another formula for memory activation which is mapped onto probability of recall similar to how ACT-R activation is mapped onto probability of recall (see later discussion leading to Equation 7). They express the central PPE equation for activation of item i as

\[ M_i = N_i^c \times T_i^{-d_i} \]

where \( N_i \) is the number of times the item has occurred, \( T_i \) is elapsed time, and \( d_i \) is the decay rate for i. Adding a scaling parameter A to convert \( M_i \) to odds of occurring in the
environment results in the following equation:

$$\text{Odds}_i = A \times N_i^c \times T_i^{-d_i} \quad (4)$$

Walsh et al. elaborate PPE in terms of the calculation of elapsed time $T_i$ and the decay rate $d_i$. They propose that the elapsed time is a weighted average of the times $t_j$ that an item has occurred:

$$T_i = \sum_{j=1}^{N_i} w_j \times t_j \quad \text{where} \quad w_j = \frac{t_j^{-x}}{\sum_{j=1}^{N_i} t_j^{-x}} \quad (4a)$$

As $x$ increases the most recent time dominates the weighted average (when $x = 1$, this calculates the harmonic mean). As $x$ gets very large $T_i$ becomes the time of the most recent time and PPE becomes no different than GPE in this regard. The lags, $\log_j$, between successive presentations determine the value of the decay $d_i$:

$$d_i = b + \frac{m}{N_i - 1} \times \sum_{j=1}^{N_i-1} \frac{1}{\log(\log_j) + e} \quad (4b)$$

When $x$ is very large and $m = 0$, the predictions of PPE become identical to those of GPE.

Like P&A, PPE calculates a new decay on each presentation of the item. However, unlike P&A or standard ACT-R, it does not require adding up the effects of multiple presentations. There is, in effect, a single trace that is decaying as a function of elapsed time. Walsh et al. (2018) emphasize the computational advantages of PPE when the number of presentations is large.

Figures 8a and b show the predictions of this model. It fits somewhat better than GPE. PPE produces spacing effects, if weaker than in the data. It could produce larger spacing effects with larger values of the $m$ parameter but this would be at cost to its overall fit to the data. In particular, higher frequencies tend to become worse than lower frequencies at long lags. For instance, the twice-occurring items in Figure 8b become worse than the once occurring items.

3.4. MCM

The other spacing model is the Multiscale Context Model (henceforth MCM) described by Mozer et al. (2009). According to MCM, memory for an item is stored as a set of $N$ PPE maps $M_i$ to recall probability by the same equation as ACT-R uses to map activation into recall probability. Given that activation in ACT-R is interpreted as log odds, we did try treating $M_i$ as log odds but this did not fit as well.

We thank Mike Mozer for his help in achieving an implementation of this model.
Figure 8: Predictions of two memory models that produce spacing effects: (a) and (b) are from PPE; (c) and (d) are from MCM.

decaying traces and the probability of recall is determined by the weighted strengths of these traces. Adding a scale parameter, $A$, to transform their predictions of probability of recall into odds of occurring in the environment, their equation becomes:

$$Odds = A \times \frac{\sum_{i=1}^{N} \gamma_i x_i}{1 - \sum_{i=1}^{N} \gamma_i x_i}$$  \hspace{1cm} (5)$$

where $x_i$ is the strength and $\gamma_i$ is the weight for trace $i$. The traces decay according to the rule

$$x_i(t + \Delta t) = x_i(t) \times e^{-\Delta t/\tau_i}$$  \hspace{1cm} (5a)$$

The $x_i$'s all start at 1 and decay according to their rates. When the item is presented again the individual traces are incremented by adding to the ith trace an amount that is a function of the average strength of the first i traces:

$$\Delta x_i = 1 - \frac{\sum_{k=1}^{i} \gamma_k x_k}{\sum_{k=1}^{i} \gamma_k}$$  \hspace{1cm} (5b)$$

The parametrization of the model concerns the determination of the decay rates $\tau_i$ and the relative weights of the traces $\gamma_i$. They propose that the decay parameters are monotoni-
cally increasing, resulting in slower decays, and the weights are monotonically decreasing, resulting in less emphasis on the slower decaying traces. Motivated by the potential to produce a power function as a sum of weighted exponentials over a large ranges of magnitudes they have used \( N = 100 \). The decay parameters and weights are power functions of \( i \):

\[
\tau_i = \mu \times \nu^i \quad (5c)
\]

\[
\gamma_i = \frac{\omega \times \xi^i}{\sum_{j=1}^{N} \xi^i} \quad (5d)
\]

With the constraints that \( \nu > 1 \) and \( \omega < 1 \) to produce the desired ordering. Figures 8c and d show the predictions of the MCM model. The quality of its fit is similar to GPE and PPE (Table 1). Like PPE it captures some of the effects of spacing as well as frequency and recency, but there are discrepancies. Perhaps most glaring is that it predicts high frequencies will be worse than low frequencies at longer delays. The massing of the items will build up high strength of the fast decaying components which will be lost at longer delays. This discrepancy could also be produced by PPE, but its best-fitting parameters spared such cross-overs at the cost of underpredicting spacing effects.

3.5. Exponential A&M

Given the less than total success of the various memory models, we examined the A&M model proposed for the environment. That model had four assumptions, the first two of which come from Burrell (1985)’s model of library borrowings:

1. Different memories have different initial desirabilities, interpreted as odds of being needed\(^{10}\). These desirabilities are distributed according to a gamma distribution with shape \( v \) and scale \( b \); hence mean odds of being needed is \( v \times b \).

2. Items decay in desirability over time according to an exponential function: \( r(t) = e^{-dt} \).

3. The rate of decay, \( d \), varies for items according to an exponential distribution with mean \( \alpha \).

\(^{10}\)The Burrell model was used to predict number of borrowings in a fixed period like a year. A&M used it to predict probability of occurring in the next small unit of time. Here we will use it to predict odds in the next text.
4. There are occasional revivals where an item returns to its original desirability from which it will decay again. This is modelled as a Poisson process with rate $\beta$ and hence the mean time between revivals is $1/\beta$.

Given just the original Burrell assumptions 1 and 2, there is a closed-form solution for the estimating the odds of needing an item that has had $n$ occurrences in time $t$:

$$\lambda(n, t) = \frac{v + n}{M(t) + 1/\beta}$$

where $M(t) = \int_0^t r(x)dx$

Adding assumptions 3 and 4 allows for a much better description of the different types of memories. When the decay rate of an item is low, it is quite stable staying close to its initial desirability. When the decay rate is high and desirability is high, the item is flash-in-the-pan, that has its moment of glory and disappears. While these 4 assumptions seem reasonable qualitative features of different memories, the exact mathematic formulations are a bit arbitrary and are likely to be approximately true at best.

There is no closed-form expression for the prior odds of needing an item given all 4 assumptions. A&M came up with predictions using Monte Carlo simulations. They sampled initial odds and rates according to assumptions 1 and 3 and added in random patterns of revival according to assumption 4. At each revival the need odds returned to the original desirability and decayed according to each items decay rate. The predicted odds was then the average of the need odds calculated from these 100,000 samples. The odds calculated this way produced results that qualitatively corresponded to practice effects, retention effects, and spacing effects in human memory.

We also used Monte Carlo simulation to generate predictions for the environmental data, treating time in the A&M model as number of texts. For each simulated item we selected a desirability from a gamma distribution with parameters $v$ and $b$, a history of revivals from an exponential distribution with parameter $\beta$, and a decay rate from an exponential distribution with parameter $\alpha$. Given these choices we randomly generated a set of occurrences of an string over a sequence of 3000 texts which provides 1999 history windows of length 1000 followed by an observation window. Rather than simply observing

---

11 One difference with the assumptions of our fitting and those in Anderson & Milson [1989]: They assumed that the decay of an item began with its first presentation, which is not an unreasonable assumption in an experiment with items being presented for study. However, in these environmental domains the item has an unknown history before the 1000-text window and has been decaying for an unknown time. We model this by randomly choosing a time (according the exponential associated with the revival distribution) before the history begins and assume it has decayed from then to the beginning of the 1000-text window.
whether the item occurred in the 1001st text, we used its probability of occurring. Finally, we estimated a scale parameter A to make the resulting probabilities have the same mean as the observed probabilities. Holding the random seed constant, we estimated a set of parameters that minimized the deviations for 500,000 simulated items. Then with an estimated set of parameters, we generated 12 million simulated items which yielded over 8 billion 1000-text histories. These estimates tend to be more accurate than the data not only because there are 4 times as many observations, but also because rather than dividing observed number of occurrence of a string by number of opportunities, we average probabilities of occurring.

Figures 9a and 9b show the predictions of the A&M model. Its measures of fit (Table 1) are comparable to the more successful models considered so far. While the model does capture effects of frequency, recency, and spacing, there are differences in the shape of the delay functions, particularly apparent in Figure 9a: Whereas in the data the high frequency (\(N > 4\)) delay functions start out negatively accelerated and become positively accelerated converging to the low frequency curves, this is reversed with a final plateau above the low frequency cases. With respect to the spacing effects in part (b): whereas the lag functions for in the data cross over with short lags best at short delays and long lags best at long delays, the predictions just converge at long delays. A&M show that at certain parameterizations their model can predict a cross over in spacing effects, but the predictions here are from the best-fitting parameters. The shape of the lag functions deviate from the data in part because an exponential decay is too rapid. An item would typically have decayed to zero by the end of the 1000-text window. The probabilities plateau above zero at long delays because of revivals that return some items to their original desirability.

3.6. Power A&M

We explored a variation on the A&M model that assumed a power-law decay, which has been used to model many decay effects in human memory because of its slower decay in the long term. We kept all the other assumptions of the A&M model and simply used \(r(t) = t^{-d}\) for the decay function, assuming a similar exponential distribution on decay rates d (although d has a different meaning now). Figures 9c and 9d show the predictions of this model. This model gives a better fit than any considered so far,
Figure 9: Predictions from two versions of the A&M environmental model: (a) and (b) assume exponential decay; (c) and (d) assume power law decay.

having a considerably higher r-squared and considerably lower mean squared error (Table 1), although there are detectable differences with the data. Perhaps the most apparent difference is the lack a frequency effect for low frequencies at a delay of 1 (Frequencies 1-16 in Figure 9c and is also manifest in Figure 9d). The model also shows some erratic values for high frequencies at long lags suggesting edge cases in the Monte Carlo simulation.

3.7. AMPE

From the point of view of the A&M model of the environment, the critical factors determining odds of occurring in the next text are

1. Currency: How long ago was the most recent revival of the item.
2. Desirability: How likely is the item when it has revived.
3. Stability: How fast has the rate been decaying.

The history of the item is only of relevance in terms of the information it provides about these three factors. If memory were estimating these factors to adapt to the structure of the environment it might behave according variant of the GPE formula (Equation 1):

\[ Odds_i = \pi_i \times T_i^{-d_i} \]  

(6)
Where $\pi_i$ is the initial desirability item $i$, $T_i$ is its currency and $d_i$ is its stability. We developed a mathematical model, AMPE (A&M Performance Equation), that used heuristic estimates of these three quantities informed by our understanding of the A&M power model.

**Currency:** The estimate of currency is a variant of the PPE estimate. Since an item has the highest probability of appearing right after it has been revived, the most recent revival has a good chance of being associated with a recent appearance of the item. Therefore, like PPE the estimate of $T_i$ should be most strongly determined by the recent appearances of the item. We chose to use a simple harmonic mean of the observed times (number of intervening texts) and included a prior time $t_P$:

$$T_i = \text{harmonicMean}(t_1, \ldots, t_n, t_P) + 1 \quad (6a)$$

where $t_1$ to $t_n$ are how long ago were the $n$ occurrences of item $i$.

**Stability:** The major deviation from PPE involves the definition of decay. We examined the relationship between decay in the simulated runs of the A&M model and the statistic that determines the decay in PPE:

$$\frac{1}{N_i-1} \times \sum_{j=1}^{N_i-1} \frac{1}{\log(lag_j)+\varepsilon}$$

In the Monte Carlo simulated histories (where we can know the decay of an item) this statistic was negatively correlated with decay ($r=-.43$) whereas PPE would assume a positive relationship. The negative correlation occurs within the A&M model because items that have both high desirability and low decay will occur repeatedly at short lags. In the Monte Carlo simulations there was an even stronger negative relationship ($r = -.63$) between decay and the Range, defined as $t_n - t_1 + 1$ where $t_1$ is the first occurrence and $t_n$ is the last, measured from the beginning of the 1000-text window. Provided there is a revival close to $t_1$ and no further revivals, the Range reflects how long an item lasts before it decays away and so should provide a good estimate of decay. In the analysis for Figure 5 Range proved to be a stronger predictor of probability of occurring than

---

13If the revival times were known and one just had to estimate $\pi_i$ and $d_i$ the problem could be treated as a non-homogeneous Poisson process with a power decay. This situation has attracted a good deal of attention in studies of temporal patterns of repairs where it is sometimes called a Duane model (Duane, 1964). There are solutions for estimating the maximum likelihood estimates for the parameters $\pi_i$ and $d_i$ (Crow, 1975; Huang & Bier, 1998; Newby, 1991). Besides the problem of not knowing the revival times, to be robust these estimates would need many more revivals than typically occur in a 1000-text window (average number of revivals in the model is just over 1 in a 1000).
frequency. We defined a quantity called the "Effective Interval" which is an average of Range and a prior gap gP:

\[ M_i = \frac{\text{Range}_i + gP}{2} \]  

(6b)

To ensure that decay rates \( d_i \) were positive, we defined decay to have an inverse relationship to the effective interval:

\[ d_i = \frac{b}{M_i} \]  

(6c)

Behavioral studies of spacing typically involve a fixed number of presentations at different spacings and so spacing is confounded with Range. There is no way to orthogonally vary the three measures of number of presentations, spacing, and Range. However, in a stepwise regression of simulation data to predict decay rate from these observable quantities, the first variable to enter is Range, then number of occurrences, and finally spacing (as defined by Walsh et al. 2018). All three have negative weights. For a fixed Range and number of presentations, the PPE measure of spacing is smallest when the spacing is constant. In the A&M model this tends to be associated with a lower decay. There is a history of behavioral research comparing an expanding schedule of presentation with constant spacing, but in their review Cepeda et al. (2006) concluded that the effects are inconsistent.

**Desirability:** The closed form estimate of desirability in Burrell’s model, \((v+n)/(M(t)+1/b)\), makes desirability a ratio of frequency, \(n\), and \(M(t)\) which is like our effective interval (\(v\) and \(b\) serve as normalizing constants). In a similar if slightly simpler vein, we defined desirability as the number of observations of the item, \(n_i\), divided by the effective interval, scaled with a parameter \(a\):

\[ \pi_i = a \times \frac{n_i}{M_i} \]  

(6d)

Note that \(M_i\) has opposing effects on odds of occurring through its effect on in desirability and decay. A large value of \(M_i\) dilutes the estimate of desirability but also slows down the effect of decay.

We fit this AMPE model to the data in the same way as the other closed-form spacing models by estimating best-fitting values of its parameters, which are \(a\), \(b\), \(tP\) and \(gP\). The resulting model might not do as well as the Monte Carlo simulations because of failures in the heuristic assumptions and because the model has one fewer parameter. However, it might also do better perhaps because it is a better formulation of the underlying processes or because of better parameter estimation with its closed-form structure.

Figure 10 compares the predictions of AMPE with the data. AMPE provides a slightly better fit to the data than the Monte Carlo simulation of the power-law decay version.
of AMM or any other model (Table 1). While it comes close to capturing all the effects of frequency, recency, there are discrepancies. We believe they are due at least in part to the simple assumption that Range is the total period from the first to the last presentation. If there have been a few occurrences bunched recently and a few more bunched 900 texts ago, it seems likely that the item rapidly decayed away 900 texts ago and had a revival recently – a pattern consistent with a high decay. In contrast, the AMPE model would treat this as having a large effective interval and a low decay. We can improve the fit by making assumptions that reduce Range estimates in such cases where there is a long blank period. For instance, in Figure 10d the model is overpredicting the probability of items that occurred twice at wide spacings, treating these as slow decaying while they may well be fast-decaying items that had a revival close to the end of the 1000-text window. While we could improve fit by adding special rules for treating such cases, this move seemed too ad hoc to include in a final model.

4. Fitting Effects in Human Memory Experiments

4.1. Spacing Effects

Two clear signatures of spacing effects in the environmental data are dramatic fall-off of the high frequency curves when there has been a long period since the last occurrence

Figure 10: A comparison of the environmental data (a and b) and the AMPE model (parts c and d).
of the item (Figure 10a) and the cross-over of the retention functions for two occurrences (Figure 10b). These are deviations from simple power functions for recency and frequency and the reason for the superior fit of the AMPE model to the simple GPE model. If the environment and memory just showed power functions of time and practice, one could argue the correspondence is just a matter two unrelated, complex systems showing power law relationships. The emergence of spacing effects in both goes beyond that and argues more strongly that effects in memory are connected to effects in the patterns of appearance in the environment.

We found that existing models of spacing effects in human memory are unable to adequately capture the patterns of appearance in environmental data sets (note many of these models did not make the claim they could). These difficulties were due to fitting the complex pattern that was in the environment, particularly involving spacing effects, at a level of detail that is not available in behavioral studies of memory. Taking inspiration from the AMP environmental model, we developed an alternative mathematical model, AMPE, that does a substantially better job of capturing the environmental data. One conclusion might be that AMPE offers a better model of human memory. However, this reasoning assumes the hypothesis that human memory is adapted to the statistics of the environment. Further it depends on the assumption that our choices of environmental sources capture the statistics that shaped human memory. It would be wise to check how well AMPE can predict the results of human memory experiments on the spacing effect.

To address how the AMPE model predicts human memory we will assume that odds in the environment is related to probability of recall in the same way as in the ACT-R model. In ACT-R memory activation is supposed to reflect log odds of appearing: \( A_i = \log(Odds_i) \). A memory will be retrieved if its activation is above a threshold \( \tau \). Memory activations have momentary noise around their expected values. Assuming this noise is logistically distributed with scale parameter \( s \), ACT-R predicts that the probability of recalling a memory with activation \( A_i \) is

\[
\frac{1}{1 + e^{-\frac{\eta - \alpha_i}{s}}}.
\]

Combining the ACT-R mapping with the AMPE equation 6 for odds might seem to require estimating two further parameters, \( \tau \) and \( s \). However, the desirability scale \( a \) in Equation 6 will get absorbed in the estimate of the threshold \( \tau \) yielding the following equation for predicting probability of recall:

\[
\text{Probability}_i = \frac{1}{1 + e^{-\frac{\eta - \alpha_i}{s}}} \quad \text{where} \quad \alpha_i = \log\left(\frac{n_i \times T_i^{-d_i}}{M_i}\right).
\]
Thus, the parameters to estimate in fitting an experiment are the threshold $\eta$, the noise parameter $s$, the prior time $t_P$ involved in defining $T_i$ (Equation 6a), the prior gap $g_P$ involved in defining $M_i$ (Equation 6b), and $b$ involved in defining $d_i$ (Equation 6c).

In comparing their PPE to P&A and the SAM model (Raaijmakers 2003, Walsh et al. 2018) fit a number of data sets, all involving spacing effects. We will take advantage of their work fitting these models, especially with the SAM model, which is expensive computationally. Table 2 reproduces their results and adds the fits of AMPE and its parameter estimates. The Appendix provides a brief description of each experiment and shows the ACT-R fits. Table 2 makes a separation between three classes of experiments that require different considerations in defining what an event is for purposes of applying AMPE, which is event-based not time-based:

a **One Day.** The first set involves single-session experiments where the application of AMPE is straightforward — each test or study opportunity is another event.

b **Between Days.** The second set involves multi-day experiments where different items are studied in different patterns over days. As noted earlier, in Bahrick’s (1979) famous experiment, sessions involved training items to criterion, leaving the number and spacing of actual presentations of the item unknown. It takes fewer presentations to reach criterion on later sessions, particularly when the session is in the same day. The experiments by Cepeda et al. involved training items to criterion.

### Table 2: Fit of Models to Spacing Experiments

<table>
<thead>
<tr>
<th>(a) One Day</th>
<th>RMSE (percentage points)</th>
<th>$r^2$</th>
<th>AMPE Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPE</td>
<td>P&amp;A</td>
<td>SAM</td>
</tr>
<tr>
<td>Beg &amp; Green (1988)</td>
<td>0.0</td>
<td>0.0</td>
<td>1.8</td>
</tr>
<tr>
<td>Bregman (1976)</td>
<td>5.4</td>
<td>6.4</td>
<td>7.3</td>
</tr>
<tr>
<td>Glenberg (1976)</td>
<td>3.6</td>
<td>4.5</td>
<td>3.7</td>
</tr>
<tr>
<td>Rumelhart (1967)</td>
<td>1.9</td>
<td>1.8</td>
<td>1.9</td>
</tr>
<tr>
<td>Young (1971)</td>
<td>4.0</td>
<td>4.0</td>
<td>2.7</td>
</tr>
<tr>
<td>(b) Between Days</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrick (1971)</td>
<td>9.8</td>
<td>6.4</td>
<td>11.8</td>
</tr>
<tr>
<td>Cepeda et al (2008)</td>
<td>4.7</td>
<td>3.8</td>
<td>8.3</td>
</tr>
<tr>
<td>Cepeda et al (2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>3.2</td>
<td>3.9</td>
<td>6.5</td>
</tr>
<tr>
<td>Exp 2a</td>
<td>4.1</td>
<td>5.3</td>
<td>10.1</td>
</tr>
<tr>
<td>Exp 2b</td>
<td>4.1</td>
<td>5.6</td>
<td>8.5</td>
</tr>
<tr>
<td>(c) Mixed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P&amp;A (2005)</td>
<td>4.8</td>
<td>5.4</td>
<td>6.4</td>
</tr>
<tr>
<td>Rawson &amp; Dunlosky (2013)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp 1</td>
<td>5.6</td>
<td>3.4</td>
<td>7.8</td>
</tr>
<tr>
<td>Exp 2</td>
<td>4.8</td>
<td>3.8</td>
<td>7.3</td>
</tr>
<tr>
<td>Exp 3</td>
<td>4.4</td>
<td>4.3</td>
<td>6.5</td>
</tr>
<tr>
<td>Average</td>
<td>4.3</td>
<td>4.2</td>
<td>6.5</td>
</tr>
</tbody>
</table>
on the first session, but the subsequent sessions involved two presentations. Since the exact timing of the presentations in a session is unknown we took advantage of the near-scale-invariance of AMPE and just counted days as intervening events in determining $T_i$. In terms of representing how many presentations were in a session we used the 3-1-2 rule that worked for the Reddit data. The first study of the item counted as 3 observations (all with $T_i = 1$), a repeat on the same day counted as 1 (with $T_i = 1$), and a repeat on another day counted as 2 (with $T_i = \text{number of days}$). This probably underestimates how many times items were seen in a session in Bahrick experiment and how many times items were seen on the first day in the Cepeda experiments, but it does capture relative exposure and attention.

c Mixed. The final set involved multi-day experiments where the manipulation involves the spacing of presentations on each day, with the number and spacing of presentations known. Because the manipulation involves spacing on a day we cannot avoid the question of how to relate spacing of presentations within a day to the passage of days. P&A had treated a day as involving the same amount of relevant time as time to go through 480 items in a session. Following that lead we chose to treat a day as 500 intervening events. A whole day undoubtedly involves more experiences than 500, just as followers of a Twitter source undoubtedly had many experiences between the tweets. Interfering events must somehow be bound to a context.

In terms of quality of fit over all these experiments, Walsh et al note that PPE and P&A are basically tied. The AMPE model joins that tie. When one considers the range of experiments involved and the differences in their procedures, materials, and populations, it would be unreasonable to suppose any model could do better. So, with respect to this class of experiments the AMPE model cannot be rejected in favor of another model. This also means that the patterns in the environment, which the AMPE model largely captures, are not inconsistent with the behavioral patterns in human memory.

The parameters of the model show considerable variation across experiments. The parameters $\eta$ and $s$, which scale log odds to probability of recall, vary the least and correspond to parameters in current ACT-R that have shown similar variation (e.g. Anderson et al. [1998]), presumably reflecting differences in procedures, materials, and populations. There are some big outliers in the estimates of prior $t_P$, but the contribution of $t_P$ is minimal and the fits would be only marginally worse with a fixed $t_P$. It is more challenging how to think about the variation $b$ and $g_P$, since fits would be much worse with fixed
values for these. Across the experiments in Table 2, b and tP are strongly correlated \((r = .983)\), resulting in small variation in decay rates despite the large variation in the Range value that contributes to effective interval M (Equation 6b). Decay is determined by the ratio of b to M and M is strongly influenced by the gP estimate. Suitable choices of b and gP can keep the variations in estimated decay rates small despite wide differences in Range, which also determines the estimate of M.

4.2. Effects when Frequency is High

Studies of percent recall tend not to examine what happens in presence of very high frequencies like those observed in the environmental data because memory will be near perfect offering no discriminative information. To take an illustrative example, Pavlik & Anderson (2005) found that an item which has appeared just 8 times 3, 6, 9, 12, 15, 18, 21, and 24 events ago was always recalled in the observed cases (320 observations). The AMPE model of recall in that experiment predicts 98.2\% probability of recall. Probability of occurring again is much less than probability of being recalled. For instance, the AMPE environmental model predicts that an item which appears 8 times in the example pattern has a 3.4\% probability item of occurring next\(^{14}\).

Latency is a way to examine effects when probability of recall has effectively reached 1. Anderson & Milson (1989) introduced and justified the idea that retrieval time should reflect an inverse power function of odds of something occurring:

\[
\text{Retrieval Time} = F \times \text{Odds}^{-\text{power}}.
\]

This has been incorporated into the ACT-R latency equation for retrieval time.

The most striking aspect of the environmental data involves the high frequency cases where there has been a long time since the last occurrence of an item. The low probabilities in these cases correspond to a latency effect called the warm-up decrement. For instance, Anderson et al. (1999) gave subjects extensive practice on one day, waited a day and then gave subjects more practice. There was a large jump in the latency of the first trial of the second day from the level achieved by the end of the first day, which dropped substantially on the second trial, and was at the level of previous day by the third trial.

---

\(^{14}\)Reflecting the specificity of experimental conditions, this exact pattern never occurred in our environmental database of 2.4 billion patterns. There were 29 cases where items appeared just 8 times in the last 1000 with their first appearance being 24 from the current time and their 8th and last appearance being 3 from the current time. 7 of these appeared in the next text.
Many of the models also had difficulty accounting for the environmental patterns shown by high-frequency items after short delays. This case has been the focus of studies of Hicks law, which describes the phenomenon that choice reaction time approximately increases as log of the number of alternatives. As number of alternatives increases, the frequency of any one item during a fixed period decreases. While the factors influencing the results in such an experiment are complex on close inspection (see Proctor & Schneider, 2018 for a review), Schneider & Anderson (2011) argue that much can be explained in terms of retrieval of the individual response rules. While their explanation focused on associative strength in the ACT-R theory, there is a role for overall frequency of items. Consider, for instance, the experiment by Hale (1968) where subjects dealt with 2, 4, or 8 alternatives in a 1000 trials. This means that they would have retrieved each the responses 500 times in the 2-alternative case, 250 times in the 4-alternative case, and 125 times in the 8-alternative case. Most experiments do not have sessions of as many as 1000 trials and many are within-subject experiments in which the same subject experiences all set sizes, frequently over different experimental sessions in a single day. Still, all of these experiments involve massive experience with few alternatives and, at least locally (e.g. a session), set sizes are confounded with frequency.

We used the environmental data to examine what the implications would be for Hicks law if it latency were an inverse of Odds as argued by A&M. We looked at items whose frequency ranged from 167 to 500 in the last 1000 observations since 167 corresponds to the item frequency with 6 equiprobable items and 500 corresponds to the frequency with 2 equiprobable items. For Hicks-law data, we used the first experiment of Schneider & Anderson (2011) who varied the number of alternatives from 2 to 6. That experiment breaks the data out into cases where an item is repeated on two consecutive trials from cases where it is not. Their results are representative in finding a shallower function relating number of alternatives to reaction time for repetitions. We split the environment data into repetition trials (this would be the first point on the x-axis in Figure 10a) and non-repetitions.

Figure 11a compares latency predictions from the environment with the Schneider and Anderson data assuming the following transformation from odds to reaction time, which was proposed in A&M and is incorporated into the ACT-R latency model:

\[
Time = Intercept + Scale \times Odds^{-power}.
\]

While the Schneider and Anderson experiment has only 6 data points, the environmental data enables plotting of many more points yielding the smooth curves plotted as solid line
Figure 11: Comparison models with the effects (dashed lines) of repetition and number of alternatives in Schneider and Anderson (2011). (a) Predictions from the environment either using 1000-text windows or micro-window (solid lines) or micro-windows (dotted lines). (b-f) Prediction from memory models using their fits to 1000-text windows in the environment.
in Figure 11a. The parameters and fit is given in Table 3. While the environmental predictions capture the relative effects for repetitions and non-repetitions, their curved form with the log x-axis testifies to the fact that the environment predicts a linear relationship between number of alternatives and time.

It should be noted that Schneider and Anderson’s experiment is very far from the 1000-trial sessions of Hale. They used micro-sessions where each item is presented 6 times, creating micro-sessions of length 12, 24, or 36 depending of the set size. To investigate whether this makes a difference we looked at micro patterns in the environmental data. This amounted to

1. Focusing on windows of 12, 18, 24, 30, or 36 texts in the environmental data, rather than windows of 1000.
2. Identifying items that occurred 6 times in those windows, which would correspond to set sizes of 2, 3, 4, 5, and 6 for the different window sizes.
3. To get an effect of repetition, separating the items according to whether they occurred last in the window.
4. Calculating the empirical odds of occurrence in the next text and then converting this to a prediction of time by estimating best fitting parameters.

The resulting fit (dotted line in Figure 11a) is quite similar but somewhat better (Table 3), reflecting that these frequency effects are similar at different scales. While the fit is better, the curved lines attest to the fact that the environment still predicts a linear relationship without the log transformation.

We took each of the models fit to the environmental data in Table 1, used their predicted odds for the 1000-text windows, and estimated a best fitting intercept, scale, and power for each. Their fits are illustrated in Figure 11 and measures of fits and parameters are in Table 3. While some of the models predict linear relationships between the log number of alternatives and time, no model other than the AMPE model fits the data better than predictions taken directly from the environment. One common feature of all the other models that produce spacing effects (P&A, PPE, and MCM) is that they fail to capture even the shallow increase for repetitions. This can be also seen in the compression of their predictions at a lag of 1 for the high frequency curves (Figures 7c, 8a, and 8c).

The frequencies contributing to Figure 11 go even higher than shown in these figures up to 500 per
Table 3: Fit to Schneider and Anderson (2011)

<table>
<thead>
<tr>
<th>Measures of Fit</th>
<th>Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>$r^2$</td>
</tr>
<tr>
<td>--------------------</td>
<td></td>
</tr>
<tr>
<td>(a) Environment</td>
<td></td>
</tr>
<tr>
<td>1000- text windows</td>
<td>47.4</td>
</tr>
<tr>
<td>micro windows</td>
<td>38.6</td>
</tr>
<tr>
<td>(b) Models</td>
<td></td>
</tr>
<tr>
<td>GPE</td>
<td>93.3</td>
</tr>
<tr>
<td>ACT-R</td>
<td>53.2</td>
</tr>
<tr>
<td>P&amp;A</td>
<td>86.0</td>
</tr>
<tr>
<td>PPE</td>
<td>51.3</td>
</tr>
<tr>
<td>MCM</td>
<td>151.3</td>
</tr>
<tr>
<td>AMPE</td>
<td>27.9</td>
</tr>
</tbody>
</table>

It is curious that AMPE does a better job of predicting these data than the environmental data from which it is inspired. This reflects the fact that at short delays, the frequency bands for the environmental data are more evenly spaced than for the AMPE model (compare Figures 10a and Figure 10c). The gaps between frequency bands in the AMPE model become narrower as we would expect from a logarithmic relationship. Thus, AMPE odds but not the environmental odds predict a near linear function of the log of the number of alternatives.

5. Conclusions

The four main conclusions of this paper are (a) There is a robust regularity in how information appears in the environment; (b) The A&M model, modified to have power-law decay, captures much of that that regularity; (c) Human memory is adapted to the statistical regularities in the environment; (d)Models of human memory can be guided by trying to fit these regularities. We will elaborate of each of these conclusions:

5.1. Robust Regularity

The Twitter data and the Reddit data showed very similar patterns of string occurrences despite the difference in how they were generated. Order of appearance in Twitter is controlled by the choices of a single individual. Order of appearance in a subreddit is
a product of a large group of individuals shaped by an interaction of votes on comments and Reddit algorithms. Despite these and other differences, they hardly reflect the full range of human experience. Anderson & Schooler (1991) found similar patterns in New York Times, caregiver speech to children, and sources of email messages, although these databases were not large enough to show all the structure in the two databases used here.

Similar patterns have been found in non-text-based environments. In a study of human-to-human contact Pachur et al. (2014) once again found similar patterns. Stevens et al. (2016) found similar regularities in the encounters of chimpanzees with other chimpanzees. Schooler et al. (2000) examined the trees visited by Howler monkeys and the places (720 meter regions) visited by baboons. While these data sets are again small compared to that used here, they show similar statistics.

5.2. The Modified A&M Model

The modified A&M model assumes (a) distributions of desirability and decay rate, (b) power law decay, and (c) occasional revivals of items to their original level of desirability. Besides the overall fit in Figures 9c and 9d, there is evidence in the environment specific to each of these assumptions assumptions:

a. Distributional Assumptions: Burrell (Burrell & Cane, 1982) chose a gamma distribution for the rates because it provides the conjugate prior for a Poisson distribution and it gave results that roughly matched the distribution of library borrowings. Anderson & Milson (1989)'s assumption of an exponential distribution of decays was because it was the simplest distribution that stretched from 0 to infinity. While these exact distributional assumptions are somewhat arbitrary, they do produce a distribution of frequency of occurrences that matches up with the environment. Figure 12a shows the distribution of how often strings occur in 1000-text windows, comparing the observed distribution with the predicted frequencies of the A&M model and with the best-fitting negative binomial which is the expectation of the Burrell model. Even though the A&M model was not fit to these distributional statistics it does capture the nearly linear relation on the log-log scale that the data show. The accelerating downward trend of the Burrell model is a necessary consequence of it predicting a negative binomial distribution. In contrast, because the A&M model also has a distribution of decays, it has more items with very high or very low probabilities and thus straightens the negative binomial curve.

b. Power Law Decay. The linear relationship for frequency in Figure 12a is also the expectation of assuming a Zipf’s law relationship for relative frequency of strings, which has been observed in many domains and is probably the result of many different complex
Figure 12: Patterns in the environment that reflect the distinct assumptions of the A&M power model.
What is special about the current situation is that there are also strong effects of when the items occur, most dramatically the decay of probability. This is what motivated the assumption of decay in both the Burrell model and the A&M model. A major change from the Burrell model and the original A&M model is the assumption that this decay has the form of a power function. The success of this modified A&M model is evidence for the power assumption, but we investigated here whether we could get more direct evidence of a power decay. To get a purer reflection of the underlying decay function, we focused on items in the 1000-text window that did not occur in the first 700 texts and occurred at least once in the next 100 texts. This pattern suggests that there was a revival just before the first appearance in the hundred text window. Then we calculated probability of appearing in the 200 texts subsequent to their first appearance. A corresponding probability was obtained from Monte Carlo simulations of the A&M exponential and power models. As Figure 12b shows, the data and the power simulations have the expected near linear relationships on the log-log scale, while the exponential simulations shows the expected accelerated decrease.

**c. Revivals** A rather strong assumption of the A&M model is that when items are revived they return with their original desirability and decay rate. To investigate this, we needed windows longer than 1000 texts to obtain prima facie evidence for two distinct revivals. We restricted the analysis to sources that allowed 2000-text windows, selecting cases without occurrences of a string in positions 1-300, 701-1300, and 1701:2000, but at least one occurrence in both positions 301-500 and 1301-1500. We took these as cases where the item had a revival somewhere in positions 301 to 500, decayed away, and had another revival somewhere in positions 1301 to 1500. We counted the number of further mentions in the 200 texts after these first mention in the two windows. Figure 12c plots how number of further occurrences in the 200 strings after the first occurrence in positions 301 to 500 predicted number of occurrences after first occurrence in positions 1301 to 1500, both in the data and the A&M model. As predicted by the assumption of a return to original desirability, frequency in the first period predicts the frequency in the second period (the relationship in the other direction is almost identical). There would be no relationship if items took on new desirabilities and decay rates upon a revival. The

---

16 In the simulations of the power AM model, a revival did occur in that 100-text block 84 percent of the time compared to 9 percent of the time in prior 100-text blocks.

17 In the simulations of the power AM model, revivals did occur in each of the 200-text blocks 89 percent of the time vs 17 percent of the time in other 200-text blocks during the no-occurrence periods.
function might seem quite shallow – an item that has 12 extra appearances in the first period averages much fewer in the second period. The shallowness reflects that items that having a high number of extra occurrences in a period is likely to be an overestimate of true desirability, as confirmed by the shallow slope in the model where we know desirability and decay are constant for an item. The slope for the model is somewhat steeper than in the data, which would be consistent with items in the environment sometimes reviving with a different desirability. In the citation literature there is something called “sleeping beauties” (Garfield, 1989), where a paper suddenly acquires many more citations than would be expected from its past history (a paper "ahead of its time").

In summary, while the assumptions of the A&M Power model may not be exactly right, there is evidence that (a) The combination of a gamma distribution of desirabilities and an exponential distribution of decays produces a distribution of string frequencies like what is observed. (b) The decay process takes the form of a power function; and (c) Items undergo revivals and come back with similar properties to their original appearance.

5.3. Memory Adapted to the Environment

Although the match between the statistics that can be collected from the environment and human memory is remarkable, one can question whether there is a causal relationship going from the environment to memory. It could be that human memory is just one of those environments that shows the remarkable regularities that other environments show – i.e., this is just a case of correlation and not causation. It could be that the environments we are studying are shaped by human memory and so causation goes in the other direction. If the mirroring of memory in the environment were perfect, such causal worries do not matter for our fourth point, which is that we can use analysis of the structure of the environment to guide memory models.

Evolutionary forces should shape memory as much as possible to make more available what it more likely to be needed. Human memory has robust similarities to the memories of other mammals (Garfield, 1989) and the shaping of memory might be quite old in the evolutionary history. On the other hand, these forces might not produce a memory that perfectly reflects the statistics in modern environments. It has been questioned (Shettleworth, 2009) whether the environments Anderson & Schooler (1991) studied (and equally the environments studied in the current paper) reflect the environment in which our memory systems have evolved. The work of Schooler et al. (2000) and Stevens et al. (2016) are partial answers this challenge and suggest at least some of these patterns reflect pre-human environments. The A&M model offers a possible explanation of what is behind
the robust regularities in these environments.

5.4. Using Environmental Structure to Understand Memory

Much more data are available about environmental patterns than can ever be obtained in an experimental study of memory. Our choices of patterns to analyze in this paper were strongly influenced by the earlier choices of A&M and Anderson & Schooler (1991). There may be many other informative patterns that we have not identified and memory researchers have not imagined. Still, given our analyses and the modified A&M model, we were led to AMPE model which seems at least as successful as the competing memory models. New analyses of the environment could lead to models yet better than AMPE.

Predictive mathematical models like AMPE are needed even if predictions could in principle be derived by Monte Carlo simulations like that of the A&M model. Given the combinatorics of possible patterns, a typical experiment or real-world situation involves a combination of conditions that hardly ever will occur in the environment or in Monte Carlo simulations of the environment. It is useful to have a system of equations that can deliver a precise prediction for any situation.

The AMPE model predicts that memory availability is the result of three factors: currency, desirability, and stability. The first two are like the concepts of memory recency and strength which play a role in many models of memory. The addition of memory-specific stability is required to explain spacing and an analog of stability can be found in models that address spacing. The AMPE equation is a simpler than all the models that address spacing except PPE where it is of similar structure. Walsh et al. (2018) emphasize the ease of using PPE.

While our applications of the AMPE model have been successful, we have not been able to produce general specifications of two key aspects of the model – what constitutes an event and what how to determine the effective interval M. Availability in AMPE decays as a function of intervening events not time. Models that focus on time do not have the problem of defining what to count as an event, but they do have problems when memory does not decay as a smooth function of clock time. While we treated events as proportional to time in the fits in Table 2b, the experiments in Table 2c reveal that one cannot use this simplifying assumption in all cases. While it might be nice to assume that events are specific to particular context like an experimental session, the passage of days without such context does result in some loss of memory. Raaijmakers (2003)’s SAM model attempted to account for spacing effects and forgetting in terms of context changing over time. Despite the difficulties of that particular model, the drift of context
in some form may offer a successful alternative to either pegging memory loss to passage of time or interfering events.

The effective interval M is supposed to be an estimate of how long an item is available before it has effectively disappeared because of decay. In AMPE it is defined as an average of a prior estimate and the Range, which is the span of events over which the item was observed. In fitting the environmental data, Range was the number of texts spanning the first and last appearances of an item. We noted earlier that if there were a bunch of appearances early, a long empty gap, and a bunch of appearances late it seems wrong to treat the whole period as the Range. Defining range becomes more problematic in dealing with behavioral data where there are long gaps between sessions. In fitting the model to the data in Table 2c we summed the spans in each day but ignored the intervening days. A general use of the AMPE model requires a general definition of how events bunch together to constitute relevant periods and how to combine these into an estimate of Range.

6. Acknowledgements

Clayton Stanley is now at Google. This work was supported by the Office of Naval Research Grant N00014-21-1-2586. The data and analyses that were used to create the figures in the paper are available at XXX
References


Oxford University Press.

Evaluating the theoretic adequacy and applied potential of computational models of
the spacing effect. *Cognitive science, 42*, 644–691.


Mathematical Psychology, 8*, 58–81.
Appendix A. Experiments on the Spacing Effect

Below is a brief description of each experiment fit in Table 2. Figures A1-A9 illustrate the fits of the AMPE model. For reference we point out corresponding figures in Walsh et al. (2018) when they exist.

Begg & Green (1988) investigated memory for once-presented items at delays of 52 or 104 and a twice-presented items 52 and 104 items ago. Their results were 36%, 26%, and 62%, recall for the three conditions. With only 3 data points this is something that all of PPE, P&A, and AMPE can fit perfectly.

Bregman (1967) compared a massed condition where items were presented 16 times in a massed condition (4 times each 11 items), an intermediate condition (2 times each 11), and a spaced condition (1 time each 11). All conditions were followed with a presentation 33 items later and then after another 9 items. Figure A1 shows the classic spacing result of worse memory for spaced items at the fixed pace (test/study 1-15), then best memory after a longer delay (test/study 16), and then after relearning on another brief interval (test/study 17). Compare with Figure 3 in Walsh et al. (2018).

Glenberg (1979) compared recall at delays that varied from 2 to 64 items after lags from 1 to 40 items. He found the classic result that short delays favored short lags and long delays favored long lags (Figure A2). Recall for very short lags is depressed in all conditions – something that P&A modeled with poorer encoding. This seems not to be
used in Walsh et al.'s implementation of the P&A model and was not used in the AMPE model. As consequence AMPE captures only the reduced effect of delay at long lags and not the depression in recall at short lags for the brief delays. It nonetheless fares comparably to the other models.

**Rumelhart (1967)** either presented items 5 times at lags of 2, 4, 7, or 11 (pure conditions) or at a mixture of lags (mixed conditions). All conditions had a final test at a lag of 11. Subjects were forced to guess from one of three alternatives on each trial and this is the reason they are about 1/3 correct on the first presentation (this was the first study opportunity). Therefore, in modeling these data we included a simple guessing model which had a 1/3 chance of being correct if the item was not recalled (Figure A3).

**Young (1971)** compared twice-presented items at various lags from 1 to 18 items then tested at a delay of 11 items and once-presented items tested at delays from 1 to 11 and the results are given in Figure A4. There might be a peak for the twice-presented items around 7-9, which would be consistent the idea that retention at an interval does best when the spacing matches that interval. Both Raaijmakers (2003) and P&A fit only the twice-presented items and produced a matching peak. While AMPE would also produce a peak if fit to just the twice-presented data, it does not if the once-presented items are also included and is the data used in Walsh et al (2018). Despite this, its fit is better.
Figure Appendix A.3: Results (a) and AMPE predictions (b) for Rumelhart (1967): Growth in recognition accuracy for items studied at various lags.
Figure Appendix A.4: Results (a) and AMPE predictions (b) for Young (1971): Recall of items presented once or twice at various lags.

than those reported by Walsh et al. (2018) for PPE or P&A. 

Bahrick (1979) reports an experiment where items were presented either 3 times or 6 times either all in one day, at 1-day separation, or at 30 days separation and then tested at a 30-day retention interval. Bahrick & Phelphs (1987) then reported a further test of the 6-repetition items 8 years later. The results generally show the superiority of 30 day spacing for long-term retention (Figure A5). Compare with Figure 11 in Walsh et al. (2018) 

Cepeda et al. (2008) varied the lag between learning sessions from 0 to 105 days, and the delay to final test (RI) from 7 to 350 days. Their results can be taken as identifying an optimal spacing for different retention intervals (Figure A6). Compare with Figure 5 in Walsh et al. (2018) 

Cepeda et al. (2009) reported 2 experiments involving 3 sessions where the first two offered study opportunities at different lags and the third session was a retention test at different at a delay of 10 days in experiment 1 and 6 months in experiment 2. The data and predictions for tests in the second session are shown in parts a and b of Figure A7. The data and predictions for the final retention test in Session 3 are in parts c and d. Compare with Figure 4 in Walsh et al. (2018) 

Pavlik & Anderson (2005) had subjects perform test-and-study trials varying num-
Figure Appendix A.5: Results (a and c) and AMPE predictions (b and d) for Bahrick (1979) and Bahrick & Phelphs (1987). Parts (a) and (b) are for items studied three times at various lags and then tested at a lag of 30. Parts (c) and (d) are for items studied 6 times at various lags and then tested at a lag of 30 and again at a lag of 8 years.
Figure Appendix A.6: Results (a) and AMPE predictions (b) for Cepeda et al. (2008): Retention results at various delays for items studied twice at various lags.
Figure Appendix A.7: Results (a and c) and AMPE predictions (b and d) for Cepeda et al. (2009): Parts (a) and (b) are for the second study session and parts (c) and (d) are for the third retention test.

Figure Appendix A.8: Results (a) and AMPE predictions (b) for Pavlik & Anderson (2005): Initial learning of items in the first session at point starts at x=1. The relearning of items in the second session is plotted starting at x=4 if there were 2 tests in session 1, at x=6 if there were 4 tests in session 1, and at x=10 if there were 8 tests.
Rawson & Dunlosky (2013) reported 3 experiments examining the effects of different lags within a day. Figure A9 shows the data for the first and third experiments, which have the most conditions. In experiment 1, subjects studied the items twice at 0,
1, 3, or 7 intervening items on a day 1 and then studied twice at lags of 7 on days 3, 8, and 10. In Experiment 3, subjects studied items at 0 and 7 lag on day 1 and crossed with that at 0 and 7 lag on subsequent days. Compare with Figure 8 in [Walsh et al.](2018).