The Environmental Basis of Memory

John R. Anderson[a,*], Shawn Betts[a], Michael D. Byrne[b], Lael J. Schooler[c], Clayton Stanley[d]

[a] *Department of Psychology, Carnegie Mellon, United States*
[b] *Department of Psychological Sciences, Rice University, United States*
[c] *Department of Psychology, Syracuse University, United States*
[d] *Unaffiliated*

Abstract

Memory should make more available things that are more likely to be needed. Across multiple environmental domains it has been shown that such a system would match observed memory effects involving repetition, delay, and spacing (Schooler & Anderson, 2016). To obtain data of sufficient size to study how detailed patterns of past appearance predict probability of being needed again, we examined the patterns with which words appear in large two data sets: tweets from popular sources and comments on popular subreddits. The two data sets show remarkably similar statistics, which are also consistent with earlier, smaller studies of environmental statistics. None of a candidate set of mathematical models of memory do well at predicting the observed patterns in these environments. A new model of human memory based on the environmental model proposed by Anderson & Milson (1989) did better at predicting the environmental data and a wide range of behavioral studies that measure memory availability by probability of recall and speed of retrieval. A critical variable in this model was range, the span of time over which an item occurs, which was discovered in mining the environmental data. These results suggest that theories of memory can be guided by mining of the statistical structure of the environment.

Keywords:
Memory, Rational Analysis, Environmental Statistics, Spacing Effect

*Corresponding author at: 5000 Forbes Ave., Pittsburgh, PA 15213
Email address: `ja@andrew.cmu.edu` (John R. Anderson)

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, 2019), have focused on how computational limitations shape cognition in the spirit of ecological rationality (Todd & Gigerenzer, 2012).

As one earlier example of rational analysis, Anderson & Milson (1989), (henceforth A&M), proposed that the goal of memory was to make most available what was most likely to be needed in the current situation. They proposed a model of how the probability of an item varied (details later in this paper) and with that model were able to predict many of the major effects in human memory such learning functions, forgetting functions, and the spacing effect. The only computational assumptions they needed were that past experiences are considered in order of their probability of being needed and that it takes more time to consider more experiences.

Simon (1989) made an early and cutting criticism of this rational analysis, particularly with its emphasis on the environment and disregard for mechanism. For instance, he commented on the A&M assumption that the need to remember an experience decayed with time. He asserted that this assumption was "motivated by our knowledge that empirically there is memory decay". He argued that rational analysis was just making up a world in which human memory would be optimal and that the assumptions in that model should be tested as to whether they were true of the actual world. In his thesis, Schooler (1993) performed just such empirical tests. 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 probable would display behavior that matched human memory.

The early rational analysis had argued for a focus on the structure of the environment instead of the structure of mind. That competition between approaches was mistaken. A careful analysis of the demands of the environment can help shape a mechanistic theory of the mind that responds to those demands. The A&M theory and the Schooler analyses guided the mechanistic theory of ACT-R (Anderson & Lebiere, 1998), which also incorporated other results from rational analysis in its design.

This paper follows up on Schooler's work by exploring much larger environmental databases. We will show that a detailed understanding of the structure of such data can lead to a more successful theory of human memory. This paper has the following 6 parts:

1. A review of Schooler's analyses which serve as the template for the new analyses reported here.
2. An examination of the detailed structure of two large databases, one involving Twitter messages and the other involving Reddit messages.
3. Tests showing that relevant theories of memory do not correspond to this structure.
4. A new theory of memory, AMPE based on the original A&M theory, that does correspond to the environmental structure.
5. Tests showing that AMPE fits behavioral data on human memory better than alternative theories.
6. Discussion of the robustness of the statistical structure of the environment and its implications for understanding human memory.

1. The Schooler Analyses

To understand the approach that Schooler took to relating the environment to memory we will review his analysis of patterns of word usage in the New York Times during the 1980s. The underlying assumption was that when a word like "Qaddafi" appeared in a headline it was a demand on the reader's memory to retrieve who Qaddafi is in order to judge whether they wanted to read the article. An adaptive memory would make knowledge available about Qaddafi to the degree that it was likely to be needed.

Anderson & Schooler (1991) examined how the pattern of word use in the last 100 days of the New York Times predicted its use on the next (101st) day. Figure 1 illustrates some of the effects they found. Figures 1a and 1b illustrate that the probability of a word appearing in a New York Times headline on the next day decreases 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 axes are log transformed – the signature of a power function. Human forgetting has also been characterized (e.g. Wickelgren, 1974; Wixted & Ebbesen,

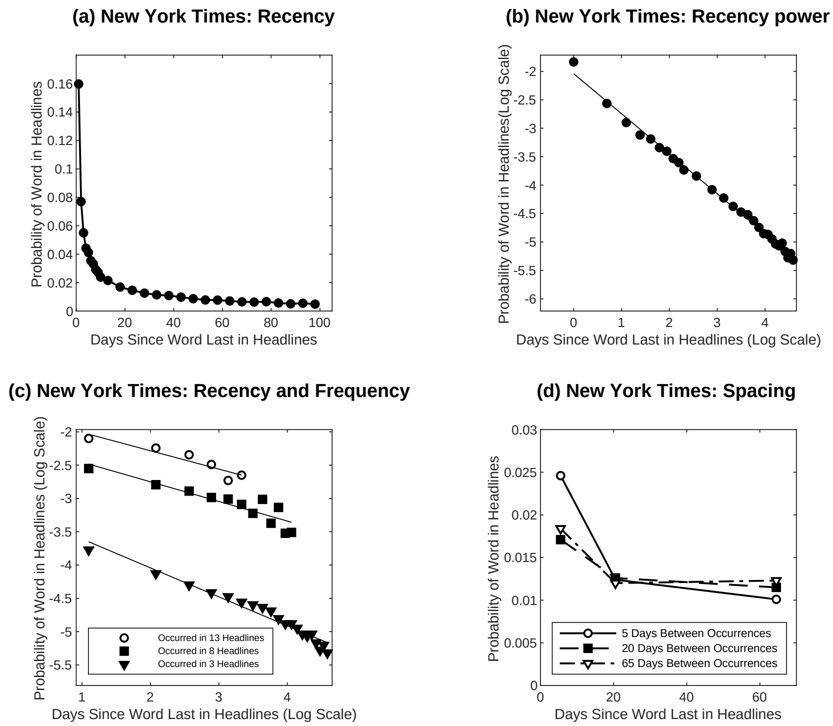

Figure 1: Anderson and Schooler's (1991) predictive patterns in the New York Times: (a) Probability of a word occurring on the next day as function of how long ago it last occurred; (b) Log-log transform of part of part a; (c) Joint effect of frequency of occurrence in the last 100 days and delay since last occurrence; (d) For items that occurred twice in the last hundred days, the effect of lag between the two occurrences and time since last occurrence.

1991) as a power function – probability of recalling a memory decreases as a power function of the amount of time that has passed. Note that here is the response to Simon's challenge to see if such a decay occurred in the real world.

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[1] are approximately a product of a power function of practice and a power function of delay. In a survey of the literature, Schooler and Anderson found that frequency and recency combined in a similar way.

Figure 1d is for words that appeared just twice in the last 100 days and shows the

[1] Odds is used in this equation because it is not bounded above by 1. There is little difference between odds and probability for the low probabilities in Figure 1.

interaction between the spacing of the two occurrences and the delay since the last occurrence. The different curves reflect different spacings between the two occurrences. Reflecting the classic spacing effect (Cepeda et al., 2006) 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 similar analyses, it is important to understand how the data in displays like Figure 1 is calculated. These analyses examine how measures calculated from the appearances of a word like ”Qaddafi” in the last 100 days of New York Times headlines predicts the probability it occurred on the 101st day. The data for these analyses come from 730 days of headlines that that span 1987 and 1988. To perform these analyses, 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 parts a and b of Figure 1, many of the points in part c, and all of the points in part d.

Anderson & Schooler (1991) found strikingly similar patterns in caregiver speech to children and in who sent JA email messages. Subsequently, similar patterns have been found in what trees Howler monkeys visit, what locations baboons visit, what other chimpanzees a chimpanzee encounters in a day, and what other people a German student encounters in a day (Pachur et al., 2014; Schooler et al., 2000; Stevens et al., 2016). This supports that conclusion that there are robust regularities in way information appears in the human environment that may reflect patterns that hold even in pre-human history. An adaptive memory would be shaped to reflect such regularities.

One limitation of the Anderson & Schooler (1991) analyses and of the later analyses is the size of the databases that were examined. This reflected both limitations in computational power in some cases and in the amount of data that was available in other cases. This means that the ability to identify the patterns that appear in Figures 1c and 1d is limited. Moreover, these analyses have not extended to temporal patterns beyond 100 (x-axis in Figure 1). Increasing that scale can reveal much about such sequential data. Behavioral experiments on memory are even more limited in the detail with which they can examine such patterns, but it is no longer necessary to similarly restrict studies of environmental demands on memory. This raises the possibility that, if the rational thesis is correct, we will be able to discover things about memory by studying the environment.

This paper is based on two much larger data sets, one from Twitter and one from Reddit, that allow for a thorough examination of the interaction between recency, frequency, and spacing (as well as other factors). Looking at the patterns that emerge from these data and performing detailed tests, we improved the A&M theory of the structure of environmental demands on memory. This then led to a new theory of memory, AMPE, that addresses the interactions of recency, frequency, and spacing more successfully than existing theories.

2. Patterns in Twitter Messages and Reddit Comments

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, where each text consists of a set of words. 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 predict its probability of occurring in the next text. Because the appearance of a string is a demand on the memory of the follower to understand the referent of the string, an adaptive memory would make most available the most probable strings. 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 database 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 7, 2014). 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.

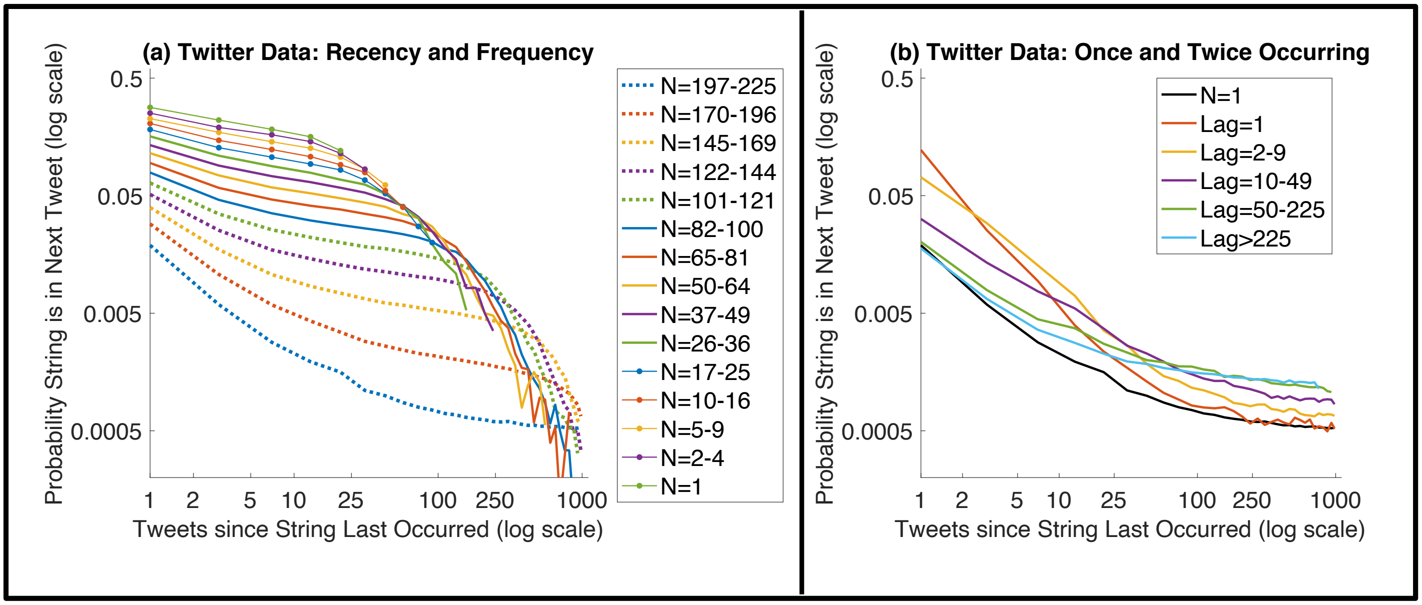

Figure 2: A strings pattern of appearance in the last 1000 tweets predicts its probability of occurring in the next tweet. The x-axis in both graphs is the number of tweets since the string last occurred. Part (a) looks at how this variable interacts with N, the number of occurrences of the string in the last 1000 tweets. Part (b) focuses on strings that occurred just once (N=1) or twice in the last 1000 tweets. For the strings that occurred twice it shows the effect of Lag, the number of tweets intervening between the two occurrence.

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. An adaptive memory should make these words available to the degree they are likely to occur. We stripped high frequency function words from the tweets, then limited all tweets to a vocabulary of the 20,000 most frequent strings (tweets do contain strings that are not words), and eliminated repeated strings in a tweet. Hashtags survived if they had high enough frequency. The surviving strings from the example tweet are:

> ’debate’ ’just’ ’numbers’ ’set’ ’major’ ’decisions’ ’going’ ’affect’ ’millions’ ’families’

This process yielded 1,038,632 tweets averaging 7.5 unique strings.[2]

Using the same moving window technique described with respect to Figure 1, we examined how the pattern of appearance over 1000 tweets predicted the probability of

[2]While frequency of retweeting varies from tweeter to tweeter, approximately one-fifth of these tweets are retweets.

occurring in the 1001st tweet. Of the top 500 English tweeters, 361 had in excess of 1000 tweets and they averaged approximately 2722 tweets[3] . To apply this analysis we need at least 1001 tweets - first 1000 are the predictive context and last is the tweet to be predicted. Thus, there are approximately 361*(2722-1001) = 621,281 windows for analysis. Each string 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. The windows average approximately 1900 distinct strings. Thus, there are approximately 621,281 * 1900 or ~1.2 billion patterns for analysis.

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.[4] To better reveal what is happening, both the x and y axes are plotted in log units. The patterns in Figure 2a correspond somewhat to what Anderson & Schooler (1991) reported – parallel linear functions of delay (see Figure 1c). 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.

While this equation captures some of the effects, there are deviations from its simple linear combination: higher frequencies show somewhat shallower slopes initially. However, at a certain point the higher bands show a sharp drop with delay to the point where differences among frequencies seem to almost disappear. This final drop off can

[3] Because our analyses are of 1001 tweet windows, it is actually irrelevant whether the reader reads all the tweets from a source or just a subsequence that involves these 1000 tweets.

[4] The nth frequency band includes $2n - 1$ specific frequencies. We similarly aggregated delays into bands that had $2n - 1$ delays and plotted these at the x-values that were the average delays for those bands. This means, for instance, that in Figure 2a 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 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 3rd curve at x=13 which is the mean of 10-16. Cases are plotted only if they 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.

be characterized as reflecting a spacing effect: If all the many presentations were massed together a long time ago, they show accelerated forgetting and are not much more likely to be remembered than a single presentation that long ago. Restricted to the lowest four frequency bands and delays less than 100, the pattern is similar to Figure 1c, which covers this range[5]. 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 long lags are better. For all of the lag bands, the effect of delay is negatively accelerated, which is 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 the number of intervening tweets. These tweets are tagged for when they were sent, allowing a comparison of the predictive power of intervening time versus number of intervening tweets. This question for the environment corresponds to the perennial question in memory research: whether the passage of time promotes forgetting or whether forgetting is 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. Both are weakly correlated with frequency (order r =- .054, time r =- .027), which has a strong effect of its own. We investigated how the log of the three variables, number of occurrence in the last 1000 tweets (N), number of intervening tweets since last occurrence (I), and days since last occurrence (T), predicted probability of occurring in the next tweet. All were highly significant predictors. In a stepwise regression they entered in the order Log(N), Log(I), and Log(T). The best fitting combination was

[5]However, the frequency bands refer to number of occurrences in the last 1000 tweets, not the last 100 days as in Figure 1c.

$$\log(\text{probability}) = .0086 + .0066{*}\text{Log}(N) - .0015{*}\text{Log}(I) - .0003{*}\text{Log}(T).$$

This means that 5 days has approximately the same effect as 1 intervening tweet. Log(I) accounted for 7.8 times the variance as Log(T).

The relatively weak effect of time corresponds to the memory phenomenon that forgetting is much less rapid outside of the context of an experiment. For instance, Anderson et al. (1999) found that the passage on a day outside of the context of a memory experiment was equivalent to approximately 10 minutes in an experiment (see also McBride & Dosher, 1997; Wickelgren, 1972). The fact that memory forgetting is more strongly driven by the passage of time in a particular experimental context than by the passage of time in general can be seen as an adaptation to what predicts the likelihood of an event occurring in the environment.

2.2. Reddit Database

Reddit is a discussion website organized by topics into subreddits. Within a subreddit there are discussions which begins with an initial posting followed by a series of hierarchically organized comments. Order of discussions and order of comments within a discussion are determined by user votes. These discussions provide a test of whether the same patterns will emerge in a sequence of texts that is produced in a quite different way. We analyzed comments 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. This contrast is important because one might argue that the memory-environment correspondence in the case of Twitter is because the tweeter's memory is shaping the environment, rather than the environment shaping memory. However, in the case of Reddit the environment is not controlled by any user's memory (this was also true of Schooler's analysis of senders of email to JA). To the degree that Twitter and Reddit 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.

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 function 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 slightly more than 1.2 billion patterns, similar to the Twitter database.

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 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 3a displays the frequency and recency effects to compare with Figure 2a. Figure 3b displays the spacing effects to compare with Figure 2b. The patterns are strikingly similar to Figure 2. Parts (a) of Figures 2 and 3 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 3a are correlated .997 and in log probability they are correlated .986. Parts (b) of Figures 2 and 3 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

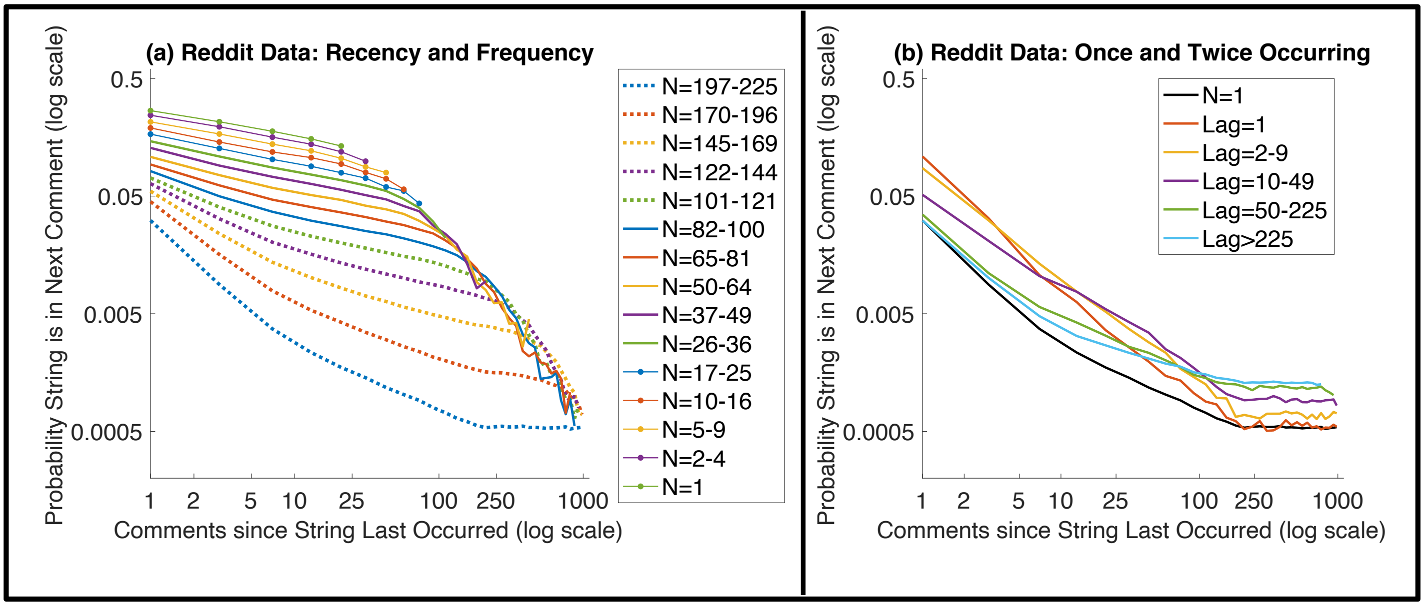

Figure 3: Predictive patterns in the Reddit data: (a) Joint effect of frequency of occurrence in the last 1000 comments and delay since last occurrence; (b) Items that occurred once or twice in the last 1000 comments – effect of lag for twice occurring.

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.

2.3. Range

In investigating the patterns in these databases we found a factor which may be more informative than spacing. This is the number of texts that span the first to last mention of a string, a variable we call **range**. Range is the sum of all the individual spacings between the pairs of mentions of a string, but theories of memory have focused on the separate effects of the individual spacings rather their total. In Figures 2b and 3b where there are just two occurrences, spacing and range are identical, but they can have different predictions when there are more than 2 occurrences.

The simplest case where range is distinct from individual spacings is when there are 3 occurrences of a string in a 1000 text window. Figures 4a and 4b show the effect of range in this case, in a form similar to Figures 2b and 3b. The interaction of range with delay can be understood the same way as the interactions in Figures 2b and 3b: Items that span longer lags or ranges are more stable. However, at short delays since the last occurrence

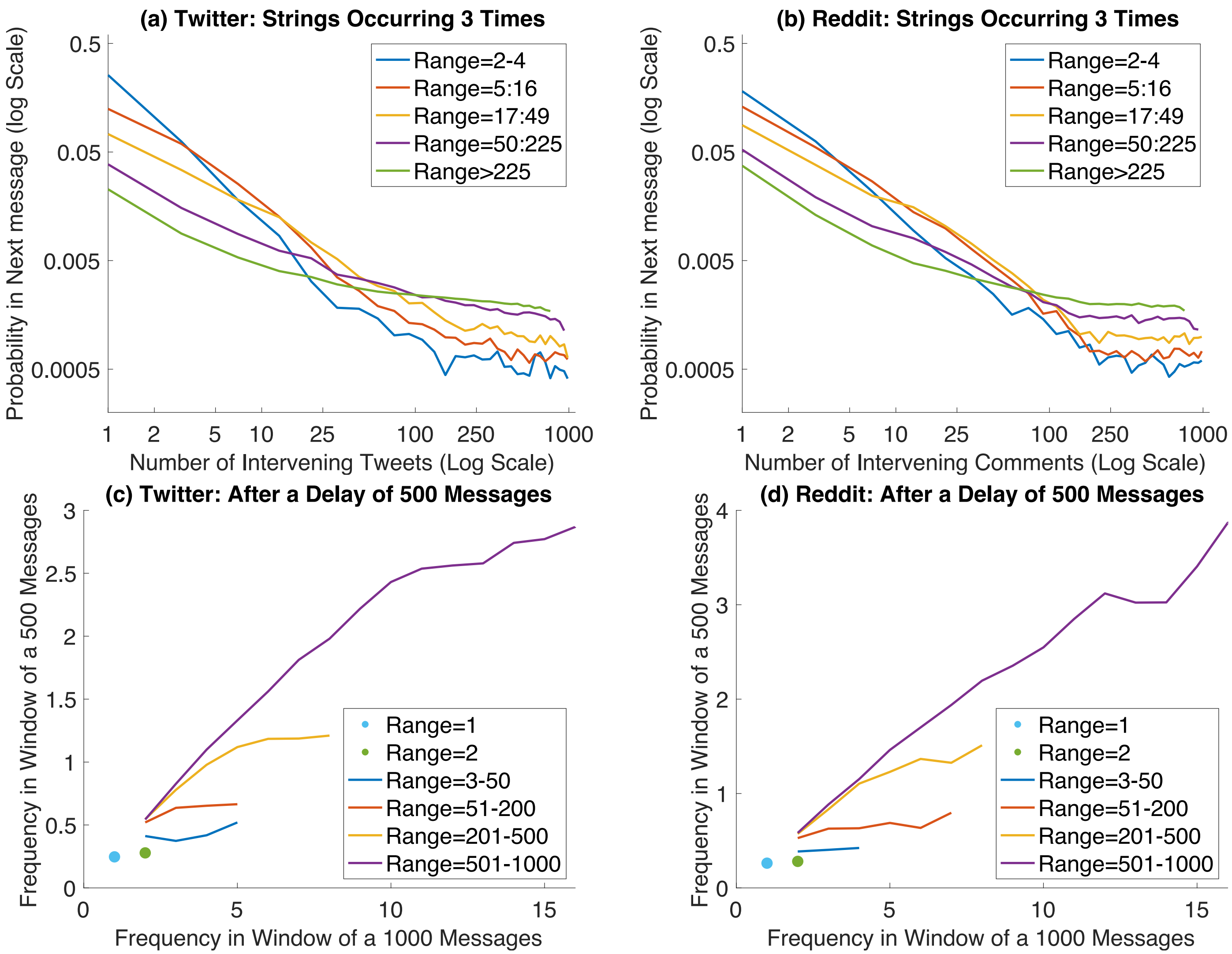

Figure 4: (a) and (b) Effect of range and delay for items that occurred 3 times in the last 1000 messages. (c) and (d) The relative effect of range and frequency in predicting the number of times an item will occur in a window of 500 messages after a delay of 500 messages without an occurrence.

(x-axis), items at short lags or ranges have the advantage that earlier occurrences are also close to the current text. As delay increases this recency advantage for the earlier items dissipates.

This leaves open the question of whether range captures all of the effect of the lags (the individual spacings) whose sum is the range. Since the effects in Figure 4a and 4b seem to stabilize after a delay of 500, we examined how the pattern of three appearances of a string predicted its frequency of occurrence after a delay of 500 texts without appearing again since its third occurrence. We counted the number of occurrences of a string an observation window consisting of 501 to 1000 texts after the third occurrence[6]. Combining Twitter and Reddit, there were 65,908 cases. The average number of occurrences in the 500-text observation window was .792 (varies from 0 to 43). Range was a much stronger predictor of this total than either of the lags and highly significant (t =17.54). After entering range as a predictor, there was no significant further variance predicted by the 2 lags (t's for entering either were .43). The difference in the BIC measures for a model using just the single predictor of range was 11.9 more than the model that used the two separate lags. We extended this comparison to cases where there were 4 occurrences in the last 1000 texts (38,887 cases) and 5 occurrences in the last 1000 texts (23,537 cases). The BIC advantage for the single variable of range versus the 3 lags in the case of 4 occurrences was 17.3 and none of the 3 lags significantly improved the prediction. The BIC advantage for the single variable of range versus the 5 lags in the case of 5 occurrences was 14.0. In this case, adding the second lag improved prediction (t = 3.80), but the other lags did not help. In summary, there may be small effects of the individual spacings, but range captures most of what has been attributed to individual spacings, at least in the case of the environment.

Parts (c) and (d) of Figure 4 compare the relatively potency of range and frequency. The x-axes give the number of occurrences of a string in a window of 1000 comments and the y-axes give the mean number of occurrence in a window of 500 messages after a delay of 500 messages without occurrence. The different points and lines reflect different ranges. An item that has a range of 1 necessarily has frequency of 1 and so it is a single

[6] More precisely for each window of 1000 texts, we focused on those strings in the 1000th text that had occurred twice earlier for a total of 3. This yielded two lags, one between the first and second and the other between the second and third. Of these strings we included only those that did not occur in the next 500 texts, guaranteeing a delay of 500. For these we counted the number of times the string occurred in the next 500 texts. Such an analysis can only be applied for sources that have at least 2000 texts.

Table 1: Fit of Various Models to the Environmental Data

	Measures of Fit		Model Parameters				
	RMSE	r^2					
(a) Mathematical Models							
GPE	0.581	0.886	c=.575	d=.608	A=.021		
ACT-R	0.825	0.784	d=.792	b=-.040			
P&A	0.684	0.847	c=.444	a=.758	B=-2.94		
PPE	0.549	0.898	x=8.699	c=.612	b=.549	m=.186	A=0.018
MCM	0.578	0.888	m=.032	u=1.111	w=.704	x=.978	A=0.029
AMPE	0.398	0.947	a=214	b=1401	tP=15.18	GP=1565	
(b) A&M Simulations of the Environment with Different Decay Functions							
Exponential	0.574	0.906	v=.164	b=0.139	a=0.035	b=333	A=0.704
Power	0.409	0.944	v=.199	b=.482	a=4.076	b=800	A=0.724

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 messages) and so it also a single point. Holding frequency constant, range has a large effect on probability of occurring. The figure also shows that frequency has a quite small effect for items that have a small range. It is almost as if all those presentations in that small range have been collapsed into a single occurrence. Overall, frequency and range have relatively equal effects. Looking at the individual strings tracked in the Twitter Figure 4c (more than 350,000 instances) number of occurrences of the strings in the target window correlates .293 with range and .230 with frequency. Looking at the individual strings tracked in the Reddit Figure 4d (almost 230,000 instances) number of occurrences correlates .295 with range and .358 with frequency. In a stepwise regression both range and frequency enter as highly significant predictors.

3. Memory Models Applied to the Environmental Data

If environmental structure shaped human memory, as the rational hypothesis implies, theories developed to account for human memory would match the environment. Anderson & Schooler (1991) showed that this was true for qualitative theories of the effects of recency, frequency, and spacing. Since 1991 there have been a number of mathematical theories of memory that address the interactions of these factors. The current data set is of such a size to provide a demanding test of how well these theories can fit the detailed patterns in the environment. The models we will review do not fare particularly well in these tests, but note that most of these models never claimed they would. Later sections will address whether this speaks against the rational analysis of memory or whether it points to weaknesses in the models because of limitations in the memory experiments that shaped these models.

The data from the Twitter Figure 2 and the Reddit Figure 3 were combined into a target data set and Table 1 examines how well various models fit those data. The first 5 models in the table are the memory models of interest. The last 3 models are models of the environment based on the original A&M (1989) proposal that we will discuss in the next section. Figure 5a displays the combined data from Twitter and Reddit for the effects of frequency and delay while Figures 5b-5f display the fits of the best fitting predictions for the 5 models. Figure 6 does the same for the effects of spacing and delay for the case of strings occurring twice in a 1000-text window. The Appendices A and B specify the details of the fitting process and the details of the models. Here we specify the basics of the fitting and explain the central equation of each model.

In fitting these and the later environmental models we minimized the sum of squared deviations from 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 data shown in parts (a) of Figures 5 and 6 are only for cases with more than 5000 observations (326 data points in Figure 5a and 187 in Figure 6a). Predictions are shown for more cases for the mathematical models because they allow for precise predictions in all possible cases. However, the fitting only involved the cases where that data had more than 5000 observations. The two measures reported in Table 1 are the residual mean square error and the R-squared.

3.1. GPE (General Performance Equation)

Anderson & Schunn (2000) proposed what they called the General Performance Equation (GPE), which applied to the current situation would make odds is the product of a power function of the number of past occurrences (N) and a power function of delay (T):

$$Odds = A \times N^c \times T^{-d} \tag{1: GPE}$$

The GPE model (Figures 5b and 6b) 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 in the GPE equation. However, it does not capture other features of the data, most notably spacing effects – the predictions are identical for all spacings in Figure 6b. While the GPE equation does not fare particularly well, variants of it are the basis for the PPE and AMPE models in Table 1, which are more successful.

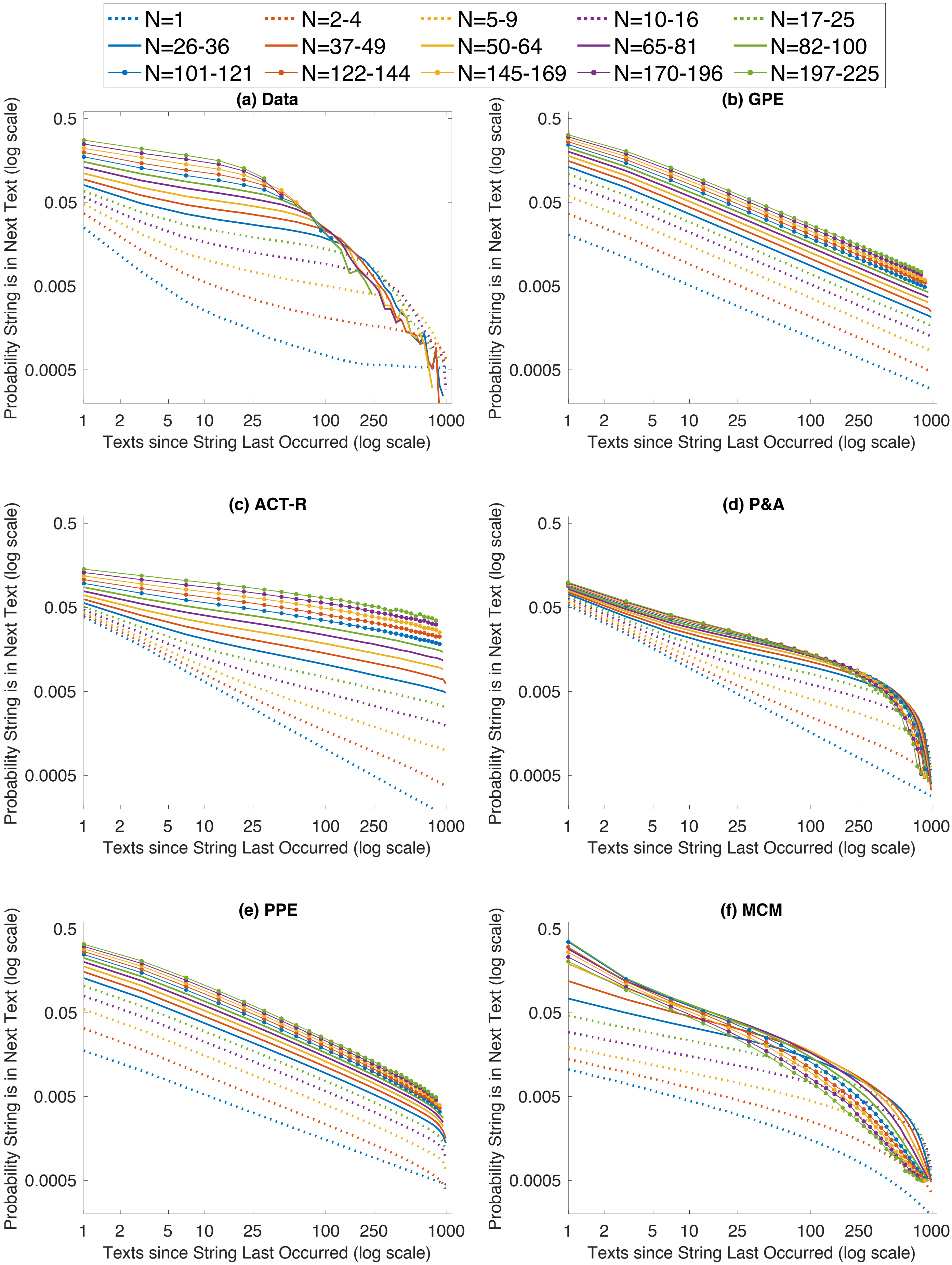

Figure 5: Effects of frequency of occurrence in the last 1000 texts and delay since last occurrence: (a) The combined Twitter and Reddit Data; (b-f) Predictions of various mathematical models of memory.

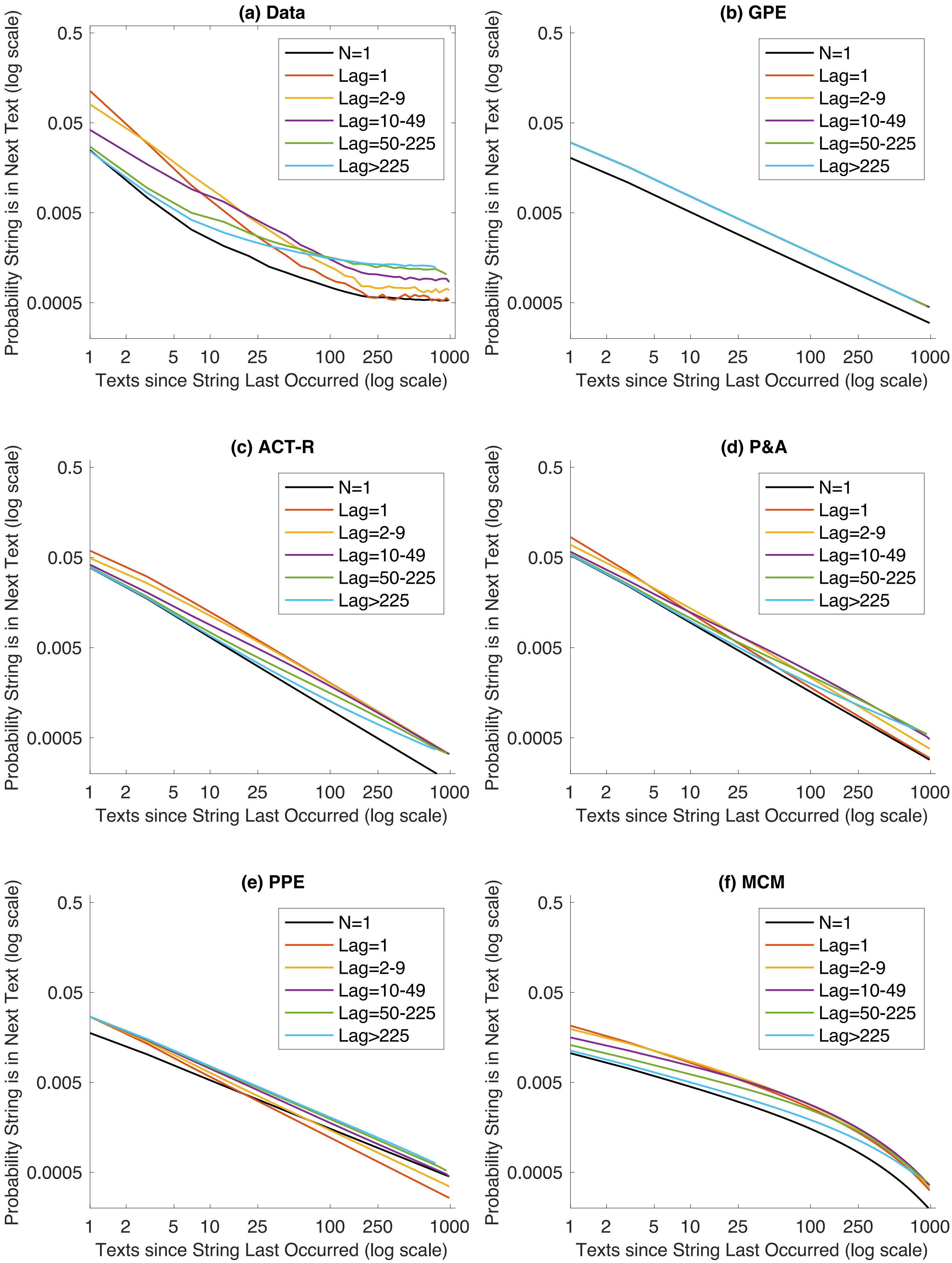

Figure 6: Strings that appeared just once (N=1) or twice in the last 1000 tweets. For the strings that occurred twice it shows the effect of Lag, the number of texts intervening between the two occurrence. (a) The combined Twitter and Reddit Data; (b-f) Predictions of various mathematical models of memory.

3.2. ACT-R Adaptive Control of Thought-Rational).

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. ACT-R uses the following equation for odds:

$$Odds = b \times \sum_{j=1}^{n} t_j^{-d} \qquad \text{(2: ACT-R)}$$

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 a parameter controlling rate of decay, and b is a scaling parameter. Figures 5c and 6c show the predictions of this model. The ACT-R model captures major trends of frequency and recency, but overall it fit is not as good as the GPE (Table 1). The delay functions in Figure 5c are 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. Figure 6c confirms the failure to capture spacing effects. The lag between the two occurrences does show an effect at short delays in Figure 6c because shorter lags mean shorter delays since the first occurrence of the item, but this advantage fades at long delays producing a convergence of the curves.

3.3. `P&A` *(Pavlik and Anderson)*

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 with higher rates in cases of close spacing (see Appendix A.1 for details). Cast as a prediction about odds (like Equation 2) their model would be:

$$Odds = b \times \sum_{j=1}^{n} t_j^{-d_j} \qquad \text{(3: PA)}$$

Figures 5d and 6d 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 these presentations are necessarily massed and therefore rapidly decaying.

3.4. PPE (Predictive Performance Equation)

Walsh et al (2008) proposed what they call the Predictive Performance Equation (henceforth PPE) as a refinement of the GPE specifically designed to address spacing effects. Applying their model to obtain a odds prediction yields

$$Odds_i = A \times N_i^c \times T_i^{-d_i} \tag{4: PPE}$$

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. The Appendix A.2 describes how spacing determines decay in their model, which is different than in P&A. They emphasize the computational advantage of their model over P&A in that one does not have to calculate a sum of decaying components. Rather their elapsed time T_i is an amalgam of the the individual t_j's, again as described in the Appendix A.2. Figures 5e and 6e show the fit of this model which is better than the previous models. PPE produces spacing effects (Figure 6e), if weaker than in the data. It could produce larger spacing effects by estimating larger d_i's for massed items, 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.

3.5. MCM (Multiscale Context Model)

Another spacing model is the Multiscale Context Model (henceforth MCM) described by Mozer et al. (2009)[7]. According to MCM, memory for an item is stored as a set of N decaying traces and the probability of recall is determined by the weighted strengths of these traces. Adding a scale parameter, A, and transforming their predictions of probability of recall into odds of occurring in the environment, their equation becomes[8]

$$Odds = A \times \frac{\sum_{i=1}^{N} \gamma_i x_i}{1 - \sum_{i=1}^{N} \gamma_i x_i} \tag{5: MCM}$$

where x_i is the strength and γ_i is the weight for trace i. The details of how the strengths decay with time and increase with practice are given in the Appendix A.3 as well as the details of how the weights are set to produce power functions. Also as described in the appendix, our implementation of the model involves a simplification because the original model becomes computationally infeasible in when frequency gets high. Figures 5f and 6f show the predictions of the MCM model. It has comparable fit measures to PPE and

[7] We thank Mike Mozer for his help in achieving an implementation of this model.

[8] To avoid infinity or other nonsense values it is necessary to bound the sum in the numerator and denominator at a value below 1 – we chose .999999. Mozer et al bound the sum at 1 for their probability predictions.

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 builds up high strength but this strength decays very rapidly. This discrepancy could also be produced by PPE, but its best-fitting parameters spared such cross-overs at the cost of underpredicting spacing effects.

3.6. Summary

All of these memory models capture the basic effects of recency and frequency, but none of them properly captures the precipitous drop for high frequency items in Figure 5a or the character of the spacing effect in Figure 6a. We have attributed the precipitous drop in Figure 5a to a spacing-like effect in that the early high-density massing of these items results in rapid decay. One might think that the failure of the spacing models (P&A, PPE, and MCM) to fit the environmental data is because spacing effects in memory are different than spacing effects in the environment. However, another possibility is that these models never addressed memory data of this density. As a first step to deciding between these possibilities we need to come to an understanding of what is happening in the environment, which is the goal of the next section.

4. Models of the Environment

As noted earlier, Anderson & Milson (1989) had derived some of the basic effects in memory from a model of how material appeared in the environment. However, that model was never compared to environmental data, as Simon (1989) noted. The current data sets offer an opportunity to determine whether their model actually describes environmental data. The A&M model started from Burrell (1985)'s model of library borrowings, which had been tested against a number of data sets. A&M had suggested that just like a library must determine how likely a book is to be borrowed to decide how available to make the book, so human memory made memories available based on an estimate of how like they were to be needed. The A&M model had four assumptions, the first two of which were identical to Burrell:

1. Different memories have different initial desirabilities, interpreted as odds of being needed. Following Burrell, 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 α, rather than the constant decay rate in Burrell.

4. Items can have spontaneous revivals (modeled as a Poisson process with rate β) where they return to their original desirability and begin to decay again. Later Burrell (2005) did consider the possibility of revival-like effects for journal articles.

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:

$$\lambda(n, t) = \frac{v+n}{M(t)+1/b} \text{ where } M(t) = \int_0^t r(x)dx$$

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 very 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. The revivals allow for such items to return. While these 4 assumptions seem reasonable qualitative features of the experiences that shape our memories, the exact mathematic formulations are somewhat arbitrary and are likely to be approximately true at best.

With these 4 assumptions, there is no closed-form expression for the odds of needing an item given a particular history of occurrences. `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.

4.1. Monte Carlo Simulations

Like Anderson & Milson (1989) 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

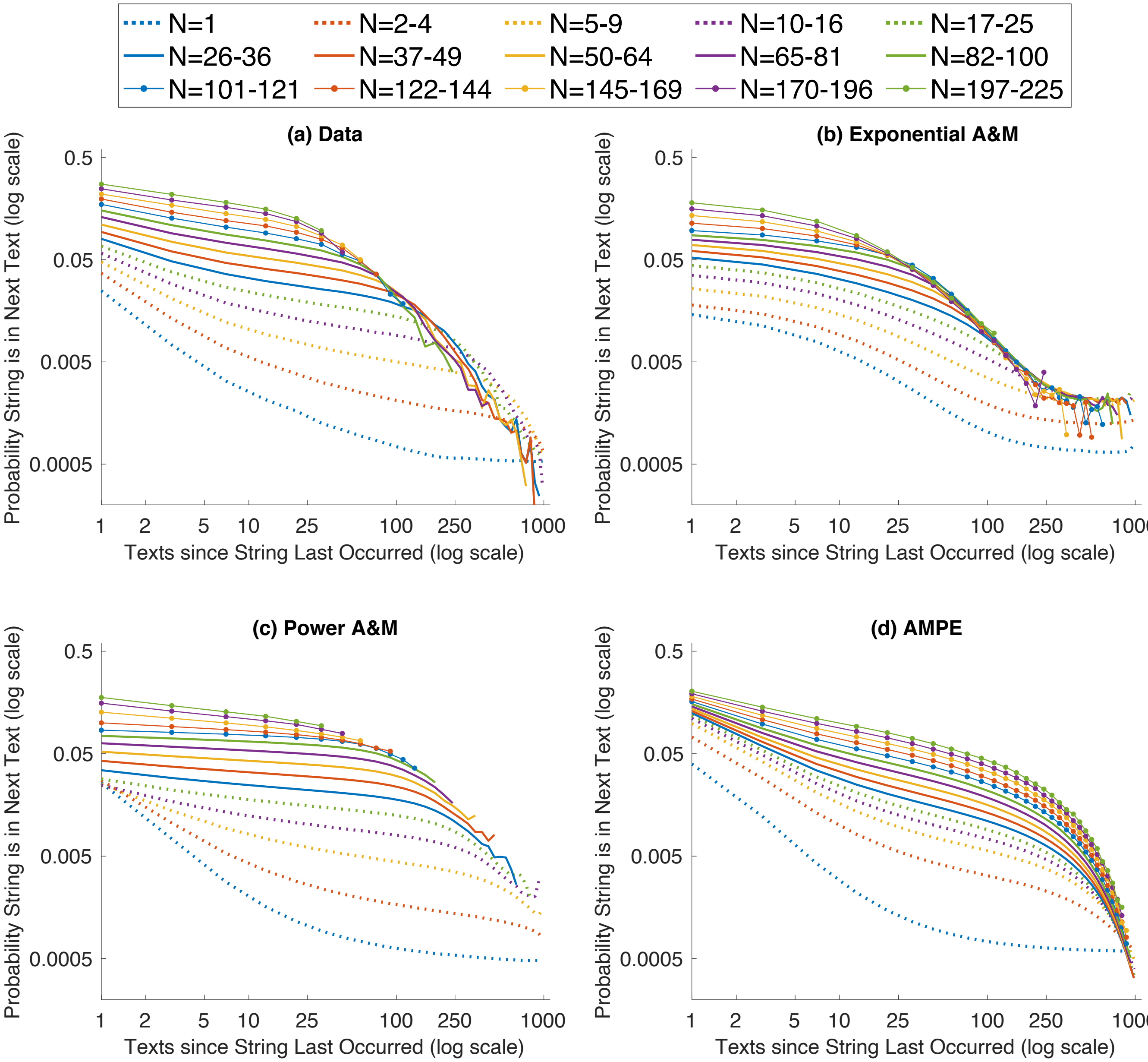

Figure 7: Effects of frequency of occurrence in the last 1000 texts and delay since last occurrence: (a) The combined Twitter and Reddit Data; (b-d) Predictions of various versions of the A&M environmental model.

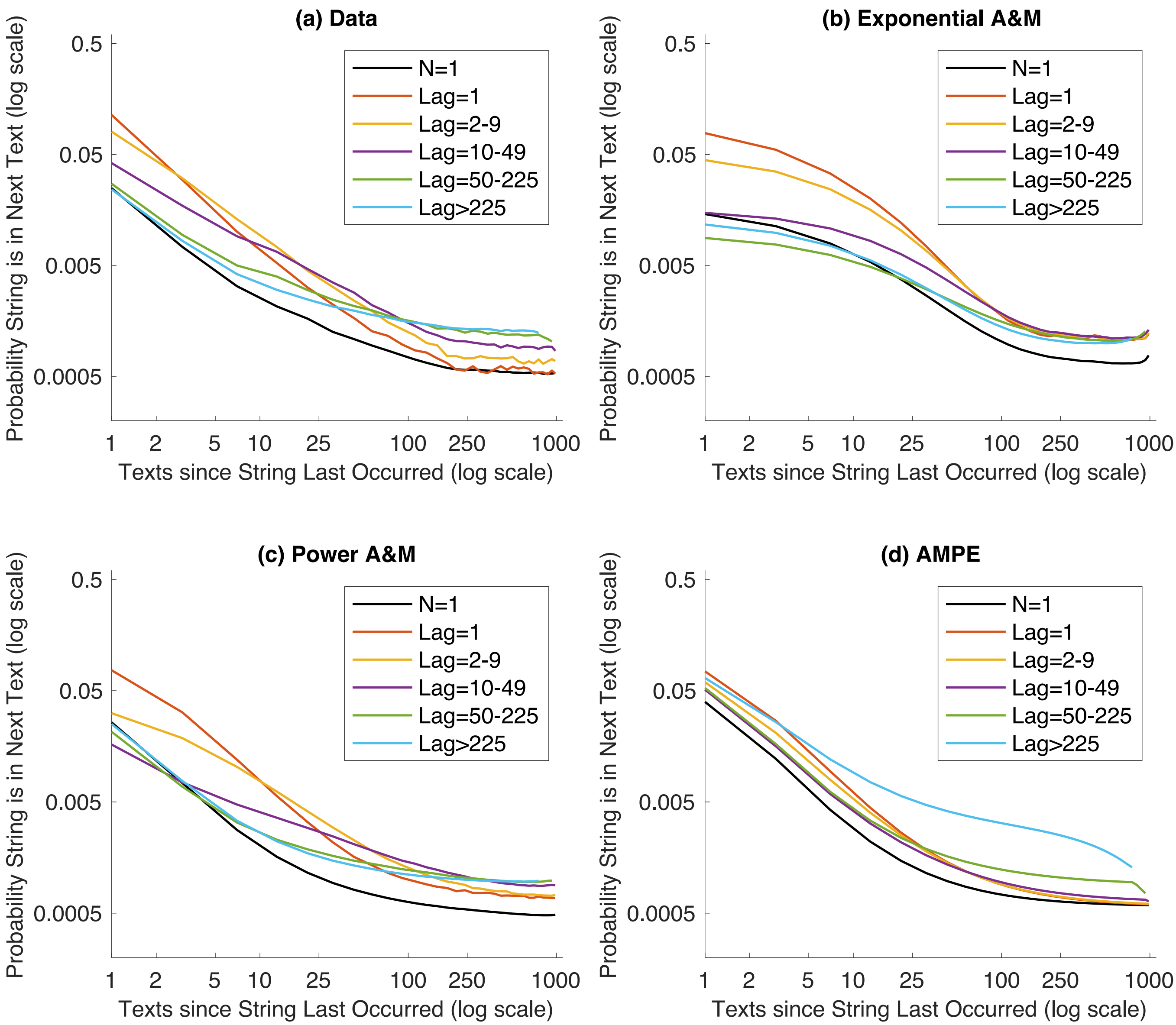

Figure 8: Strings that appeared just once (N=1) or twice in the last 1000 tweets. For the strings that occurred twice it shows the effect of Lag, the number of texts intervening between the two occurrence. (a) The combined Twitter and Reddit Data; (b-d) Predictions of various versions of the A&M environmental model.

β, and a decay rate from an exponential distribution with parameter α[9]. Given these choices we randomly generated a set of occurrences of a string over a sequence of 3000 texts which provides 1999 history windows of length 1000 texts followed by an observation text. Rather than simply observing whether the item occurred in the 1001st observation 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[10], we estimated a set of parameters that minimized the deviations for 500,000 simulated items. Then with an estimated set of parameters, we generated 18 million simulated items which yielded over 12 billion 1000-text histories[11]. The resulting predictions have less noise than the data not only because there are 5 times as many observations, but also because rather than dividing number of occurrences by number of opportunities, we average probabilities of occurring. Like the data there are cases with few observations and we kept predictions based on cases that had at least 1000 observations. There were 585 such cases, which is more than the 513 cases kept in the case of the data.

Figures 7b and 8b show the predictions of the A&M model with the combined data reproduced again in Figures 7a and 8a. 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 7b: 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 Figure 8b: whereas the lag functions for the data cross over with short lags best at short delays and long lags best at long delays,

[9] Anderson & Milson (1989) 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.

[10] This avoided the possibility that a particular parameter choice might have not a "lucky" or "unlucky" random draws. It also guaranteed that the same parameter values would yield the same results upon repetition in the search.

[11] We harvested these from 10 times more Monte Carlo 1000-text windows since the majority of the generated windows have no occurrence of the target item. This is roughly consistent with the data where only about 10% of 20,000-string vocabulary occurs in a 1000-text window.

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 delay 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 at long delays plateau above 0 because of revivals that return some items to their original desirability.

We explored a variation on the A&M model that assumed a power-law decay, which has been used to model many forgetting 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 7c and 8c show the predictions of this model. This model gives a better fit than any model considered so far, having a considerably higher R-squared and considerably lower mean squared error (Table 1), although there are detectable differences from 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 7c and is also manifest in Figure 8c).

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. Together these assumptions produce the fits in Figures 7c and 8c. To address Simon's 1989 challenge to test the model's assumptions against the environment, Appendix C reports evidence specific to each of these three assumptions.

While the A&M model with power-law decay appears to be capturing much of the structure of the environment, the need for Monte Carlo simulation limits its potential extension to a model of human memory. The process of deriving predictions is much more expensive than with a mathematical model and the predictions that are derived are only approximate. Also just as many specific patterns are rare (sometimes non-existent) in the data, they are rare in the Monte Carlo simulations. Thus, it is not practical to produce predictions for the specific patterns, in contrast to the mathematical models of Figures 5 and 6 where precise predictions could be derived for any possible pattern. This limitation is serious because most memory experiments involve collecting data for many replications of exact patterns of presentation that almost never occur in the environment or in the Monte Carlo simulations.

4.2. The AMPE Mathematical Model

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

1. Desirability(π_i): What is the desirability (initial odds of occurring) of that item?
2. Currency(T_i): How long has that item been decaying from its initial desirability?
3. Stability(d_i): How fast does the desirability decay?

The history of the item is only of relevance in terms of the information it provides about these three factors. If we had exact estimate of these quantities, a variant of the GPE Equation 1 would give the odds of an item being needed:

$$Odds_i = \pi_i \times T_i^{-d_i} \tag{6: AMPE}$$

We developed a mathematical model, AMPE (`A&M` Performance Equation), which uses heuristic estimates of these three quantities informed by our understanding of the `A&M` power model.

Currency: Since an item has the highest probability of appearing right after it has been revived, appearances of the item are hints that it has revived. Since only the most recent revival is relevant the most recent appearances of the item are more informative. We use the harmonic mean as a way of averaging the past times which emphasizes the most recent times (the harmonic mean is a special case of the averaging function used to calculate T_i in PPE):

$$T_i = harmonicMean(t_{i,1}, ..., t_{i,n}, tP) + 1 \tag{6a}$$

where $t_{i,1}$ to $t_{i,n}$ are how long ago were the n occurrences of item i and tP is a prior time added to help stabilize the estimate.

Stability: We ran Monte Carlo simulations of the power A&M model where we knew what the decay rate was and could investigate what statistic in the data predicted it. There was a strong negative relationship (r = -.63) between decay and the range, which had a strong effect in the real data (Figure 4). Range reflects how long an item stays around before it decays to a level where it is no longer available. The actual time it was available is probably somewhat longer than the observed range and so we tempered range and a prior gap gP to define a quantity M_i:

$$M_i = \frac{Range_i + gP}{2} \tag{6b}$$

To ensure that decay rates d_i were positive, we defined decay to have an inverse relationship to M_i:

$$d_i = \frac{b}{M_i} \tag{6c}$$

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 M_i (v and b serve as normalizing constants in the Burrell expression). In a similar if slightly simpler vein, we defined desirability as the number of observations of the item, n_i, divided by M_i, scaled with a parameter a:

$$\pi_i = \frac{a \times n_i}{M_i} \tag{6d}$$

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. Figures 7d and 8d show the predictions of AMPE, which provide a slightly better fit to the data than the Monte Carlo simulation of the power-law decay version of `A&M` and considerably better than any other model (Table 1). While it comes close to capturing all the effects of frequency and 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 8d 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.

5. Fitting Human Memory Data

AMPE is a mathematical model that provides a better fit to the environmental data than the other mathematical models considered. The rational-analysis hypothesis implies that it should provide a better model for human memory. This section puts that hypothesis to test. Since the best performing alternative models all addressed the spacing effect we will start out by examining its success in predicting such effects in the memory literature. Then we will consider AMPE's success in predicting memory results at high frequencies.

The AMPE model does a good job of predicting effects of frequency, recency, and spacing in the environment that are qualitatively like those that occur in human memory experiments. However, data sets from memory experiments have statistical patterns quite unlike what occurs in natural environments. This is because their presentation patterns were created to achieve particular experimental goals. For instance, an experiment might present subjects with a long sequence of items reflecting one specific spacing pattern and nothing else. Such a simple repetitive pattern would not occur in a natural environment. Thus, it is not obvious that AMPE will even produce qualitatively similar patterns when presented with such patterns, let alone quantitative fits that are competitive with other models.

5.1. Spacing Effects

To address how the AMPE model predicts human memory we will assume that odds in the environment are 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 τ. 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

$$\frac{1}{1+e^{\frac{\tau-A_i}{s}}}.$$

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

$$Probability_i = \frac{1}{1+e^{\frac{\eta-\alpha_i}{s}}} \qquad where \qquad \alpha_i = log(\frac{n_i \times T_i^{-d_i}}{M_i}). \tag{7}$$

Thus, the parameters to estimate in fitting an experiment are the threshold η, the noise parameter s, the prior time tP involved in defining T_i (Equation 6a), the prior gap gP involved in defining M_i (Equation 6b), and b involved in defining d_i (Equation 6c).

Table 2: Fit of Models to Spacing Experiments (note best-fitting parameters for Begg and Green are underdetermined)

	RMSE (percentage points)				r^2				AMPE Parameters				
(a) One Day	PPE	PA	SAM	AMPE	PPE	PA	SAM	AMPE	b	tP	gP	η	s
Begg & Green (1988)	0.0	0.0	1.8	0.0	1.00	1.00	0.99	1.00	2.2	200	9	-3.29	0.47
Bregman (1976)	5.4	6.4	7.3	5.5	0.95	0.92	0.90	0.92	320.9	9	452	-7.40	0.66
Glenberg (1976)	3.6	4.5	3.7	3.8	0.88	0.80	0.87	0.86	12.8	1000	27	-3.93	2.11
Rumelhart (1967)	1.9	1.8	1.9	2.1	0.99	0.99	0.99	0.99	82.3	2	103	-6.51	0.63
Young (1971)	4.0	4.0	2.7	3.0	0.83	0.81	0.92	0.90	177.0	2	236	-6.70	0.81
(b) Between Days													
Bahrick (1971)	9.8	6.4	11.8	9.4	0.89	0.95	0.83	0.89	178.5	998	1000	-5.80	0.41
Cepeda et al (2008)	4.7	3.8	8.3	5.3	0.97	0.98	0.92	0.97	25.1	780	184	-4.17	0.20
Cepeda et al (2009)													
Exp 1	3.2	3.9	6.5	1.7	0.96	0.94	0.83	0.99	18.5	4	78	-3.37	0.29
Exp 2a	4.1	5.3	10.1	4.4	0.97	0.95	0.82	0.97	6.4	1000	27	-3.12	0.55
Exp 2b	4.1	5.6	8.5	3.7	0.98	0.97	0.91	0.98	146.3	1000	601	-5.84	0.46
(c) Mixed													
P&A (2005)	4.8	5.4	6.4	5.6	0.97	0.96	0.93	0.95	280.9	404	752	-8.35	0.72
Rawson & Dunlowsky (2013)													
Exp 1	5.6	3.4	7.8	3.4	0.88	0.96	0.81	0.95	15.8	25	51	-4.51	0.55
Exp 2	4.8	3.8	7.3	6.2	0.91	0.94	0.78	0.81	6.0	1000	9	-5.40	2.24
Exp 3	4.4	4.3	6.5	5.1	0.94	0.95	0.88	0.92	23.3	36	56	-5.13	0.75
Average	4.3	4.2	6.5	4.2	0.93	0.94	0.89	0.94					

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 AMPE 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. In Bahrick's famous experiment, sessions involved training items to criterion, leaving the number and spacing of actual presentations of the items 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 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 a 3-1-2 rule: The first study of the item counted as 3 observations (all with $T_i = 1$), a repeat on the same day counted as 1 observation (with $T_i = 1$), and a repeat on another day counted as 2 (with T_i = 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 session 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 does a tiny bit better than either and so joins that tie. Note that AMPE considers only range while these other models consider individual spacings and these experiments were largely designed to investigate effects of individual spacings. We have shown that range and not individual spacings is critical in predicting occurrence in the environment. These results indicate it also does well in fitting the results of memory experiments that focus on individual spacings.

The parameters of the model show considerable variation across experiments. The parameters η 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 tP, but the contribution of tP is minimal and the fits would be only marginally worse with a fixed tP. It is more challenging how to think about the variation in b and gP, since fits would be much worse with fixed values for these. Decay is determined by the ratio of b to M and M is an average of range and the gP estimate (Equations 6b and 6c). Across the experiments in Table 2, b and

gP are correlated (r = .77), resulting in small variation in decay rates despite the large variation in range. The ratios of b to gP are smaller for the between-days experiments than the within-day experiments, resulting in smaller decay rates. This may reflect a slowing of decay over long time spans.

5.2. Effects when Frequency is High

There were surprising effects in the environment for high-frequency items. Studies of percent recall tend not to examine what happens in the presence of very high frequencies because memory will be near perfect offering no discriminative information. Latency is used to examine effects when probability of recall has effectively reached 1.

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 of occurring in these cases correspond to a latency effect called the **warm-up decrement**, a momentary loss of fluency after a long delay. For instance, Anderson et al. (1999) gave subjects extensive practice on one day, waited a day, and then gave subjects more practice. Their mean latency on the first recall trial of the second day for an item was 3.08 seconds longer than the level achieved by the end of the first day, but their latency on the second trial on that item sped up 2.13 seconds approaching the level of the first day. There are similar warm up effects in the environment. For instance, consider items that have occurred at least 50 times in the last 1000 but have not occurred for 100 texts. The probability of occurring in the 1001st text has dropped from their average rate of 11.70% before the hiatus to 0.97%. However, if they occur in the 1001st text their probability jumps to 8.72% in the 1002nd text. Such a jump would be expected in the AMPE model because the appearance in the 1001st text suggests a recent revival.

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 Hick's 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 over 1000 trials. This means that they would have retrieved each of 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 (within a session), frequency decreases with set size.

Set size also changes spacing in that repetitions of the same item will be closer together when set size is smaller. However, set size is nearly uncorrelated with range which is the gap between first and last presentation. Thus, this domain offers an opportunity to contrast AMPE with it focus on range with other models that focus on individual spacings.

We used the environmental data to examine what the implications would be for Hick's law if latency were an inverse of odds as argued by A&M and as incorporated into ACT-R. 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 (see Figure 9a). 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 9a) and non-repetitions.

Figure 9b compares latency predictions from the environment with the Schneider and Anderson data assuming the following ACT-R transformation from odds to reaction time, which is based on A&M's proposal:

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

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 in Figure 11a. The parameters and fit are given in Table 3.

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

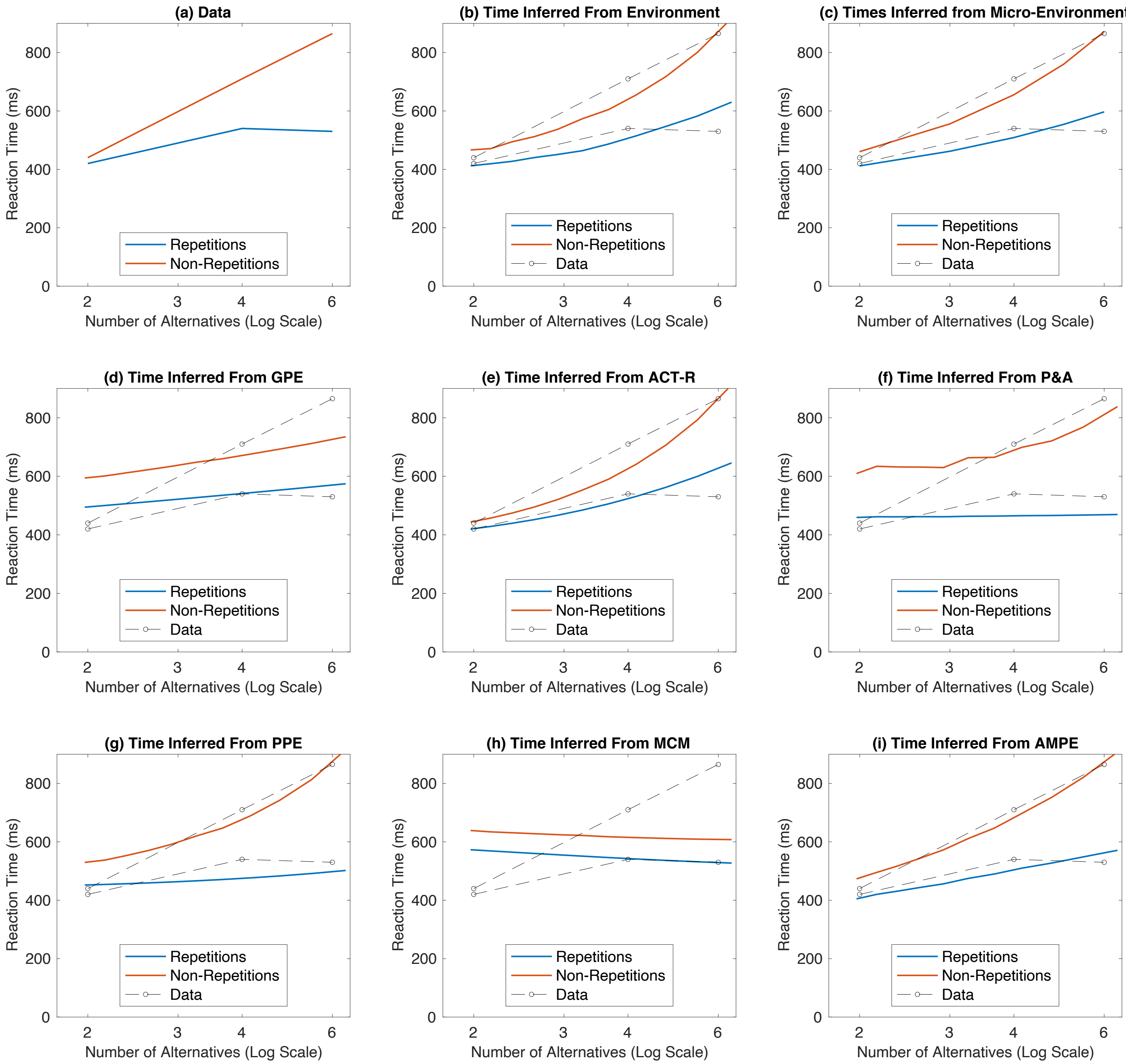

Figure 9: (a) Data from Schneider and Anderson (2010). (b)-(i) fits of memory models.

Table 3: Fit to Schneider and Anderson (2011)

	Measures of Fit		Model Parameters		
	RMSE	r^2	Intercept	Scale	Power
(a) Environment					
1000- text windows	47.4	0.908	386.1	41.4	1.40
micro windows	38.6	0.939	345.2	35.4	1.27
(b) Models					
GPE	93.3	0.826	0.0	199.8	0.22
ACT-R	53.2	0.885	349.7	13.9	1.52
P&A	0.9	0.699	450.1	0.01	3.22
PPE	51.3	0.893	430.6	16.2	1.66
MCM	151.3	0.068	0.0	433.9	0.12
AMPE	27.9	0.968	235.8	81.5	0.89

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 9c) is quite similar but somewhat better (Table 3), reflecting that these frequency effects are similar at different scales. Both predictions from the environment are somewhat bowed-shape. This is not a necessary consequence of the environmental data but a consequence of the best-fitting parameters (Intercept, Scale, and power) that map environmental odds onto latency. Either the repetition or non-repetition data could separately be fit nearly perfectly.

We took each of the mathematical 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 9d-i and measures of fits and parameters are in Table 3. The AMPE model fits the data much better any of the other mathematical models and also does a bit 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.

6. General Discussion

The idea that forgetting in memory may be adapted to the environment can be traced at least back to Renda (1910) (see Bean, 1912). Recently, a number of ideas have been advanced as to what the brain mechanisms might be that implement such adaptive forgetting (e.g., M. C. Anderson & Hulbert, 2021; Ryan & Frankland, 2022). This paper has not addressed the issue of brain mechanisms, but rather has focused on the statistical structure of the environment as the driver of this adaptive forgetting. Besides the forgetting effects of recency it has also extended the adaptive perspective to effects of frequency and range and addressed how memory availability emerges in response to the interaction of these three factors in 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 have 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 (although it is hard to make that argument for the case of Reddit). Quite possibly the causal structure is bidirectional with the environment shaping memory and our memory shaping what we put in the environment. As Lewontin (1983) puts it, organisms construct their own ecological niches. If the mirroring of memory in the environment were perfect, such causal worries would not matter for using the structure of the environment to guide memory models.

To prove that the environment causes memory one would have to manipulate the environment and look for changes in memory. If one thinks this causation occurs on a long evolutionary scale, such an experiment seems impossible. If one thought that such shaping occurred within a lifetime, such an experiment might be possible. To our knowledge the only experiment (R.B. Anderson et al., 1997) addressing this possibility did find that one could change the rate of forgetting in an experimental session by changing the likelihood with which items were encountered.

We are inclined to believe that the causation has occurred on a long evolutionary scale. 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 we are inclined to believe that 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, which would limit the relevance of environmental studies like those in this paper. 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 across all of these environments. Many environments would offer different types of events that would vary in their frequency, tend to become less frequent with time, and occasionally come back in full force.

6.1. 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, we were led to recognize the importance of range and develop the AMPE model which is at least as successful as memory models that emphasize individual spacings. 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 which is of similar structure. Walsh et al. (2018) emphasize the ease of using PPE.

6.2. Limitations of the AMPE model

While the applications of the AMPE model have been successful, we have not been able to produce truly general specifications of two key aspects of the model – what constitutes an event and how to define range. 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.

In AMPE range is the factor that determines M (Equation 6b) which is used in the definition of decay (Equation 6c) and desirability (equation 6d). 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 ranges for 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.

While there are difficulties in coming up with a fully general definition of range, we believe it is an important concept to bring into the memory literature. It may capture effects in memory better than spacing (which is equally challenging to define generally). All three of the spacing models we considered (`P&A`, PPE, and MCM) have the property that the net decay rate of an item can increase if more studies are crammed into the same interval. Depending of the situation and parameter estimates this can produce bizarre predictions. If one uses range as in AMPE, this can never happen under any parameterization. There are many demonstrations that, for purposes of long-term retention, cramming one's study over a short range is inferior to distributing it over a longer period, but to our knowledge there are no demonstrations that more study in any fixed interval is worse than less study.

7. Acknowledgements

Clayton Stanley is now at Google. This work was supported by the Office of Naval Research Grant N00014-21-1-2586. We thank Cvetomir Dimov and Christian Lebiere for their helpful comments on the work. The data and analyses that were used to create the figures in the paper are available at
`http://act-r.psy.cmu.edu/?post_type=publications&p=32939` .

References

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341–380.

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

Anderson, J. R., & Lebiere, C. J. (1998). *The atomic components of thought*. Psychology Press.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703–719.

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

Anderson, J. R., & Schunn, C. (2000). Implications of the act-r learning theory: No magic bullets. *Advances in instructional psychology, Educational design and cognitive science*, (pp. 1–33).

Anderson, M. C., & Hulbert, J. C. (2021). Active forgetting: Adaptation of memory by prefrontal control. *Annual Review of Psychology*, *72*, 1–36.

Anderson, R. B., Tweney, R. D., Rivardo, M., & Duncan, S. (1997). Need probability affects retention: A direct demonstration. *Memory & Cognition*, *25*, 867–872.

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

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

Bean, C. H. (1912). *The curve of forgetting*. Press of the New era printing Company.

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

Bregman, A. S. (1967). Distribution of practice and between-trials interference. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *21*, 1–14.

Burrell, Q. L. (1985). A note on ageing in a library circulation model. *Journal of Documentation*, *41*, 100–115.

Burrell, Q. L. (2005). Are "sleeping beauties" to be expected? *Scientometrics*, *65*, 381–389.

Burrell, Q. L., & Cane, V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society: Series A (General)*, *145*, 439–463.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246.

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

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

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive psychology*, *38*, 191–258.

Garfield, E. (1989). Delayed recognition in scientific discovery-citation frequency-analysis aids the search for case-histories. *Current Contents*, *23*, 3–9.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*, 273–278.

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

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Hale, D. (1968). The relation of correct and error responses in a serial choice reaction task. *Psychonomic Science*, *13*, 299–300.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*, 279–311.

Lewontin, R. C. (1983). Gene, organism and environment. *Evolution from molecules to men*, *273*, 975.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, 1–85.

Mandelbrot, B. (1959). A note on a class of skew distribution functions: Analysis and critique of a paper by ha simon. *Information and Control*, *2*, 90–99.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

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

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R. V., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. *Advances in neural information processing systems*, *22*, 1321–1329.

Pachur, T., Schooler, L. J., & Stevens, J. R. (2014). We'll meet again: Revealing distributional and temporal patterns of social contact. *PloS one*, *9*, e86081.

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

Proctor, R. W., & Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, *71*, 1281–1299.

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

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

Renda, A. (1910). *L'oblio: saggio sull'attività selettiva della coscienza* volume 172. Fratelli Bocca.

Rumelhart, D. E. (1967). *The effects of interpresentation intervals on performance in a continuous paired associate task (Tech. Rep. No. 116).*. Stanford, CA: Stanford University Institute for Mathematical Studies in the Social Sciences.

Ryan, T. J., & Frankland, P. W. (2022). Forgetting as a form of adaptive engram cell plasticity. *Nature Reviews Neuroscience*, (pp. 1–14).

Schneider, D. W., & Anderson, J. R. (2011). A memory-based model of hick's law. *Cognitive Psychology*, *62*, 193–222.

Schooler, L., Silva, J. C. S., & Rhine, R. (2000). Does human memory reflect the environment of early hominids? In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 22.

Schooler, L. J. (1993). Memory and the statistical structure of the environment. *Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA*, .

Schooler, L. J., & Anderson, J. R. (2016). The adaptive nature of memory. In *The Curated Reference Collection in Neuroscience and Biobehavioral Psychology* (pp. 265–278). Elsevier Science Ltd.

Shettleworth, S. J. (2009). *Cognition, evolution, and behavior*. Oxford University Press.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425–440.

Simon, H. A. (1989). Cognitive architectures and rational analysis: Comment. In *Architectures for intelligence* (pp. 37–52). Psychology Press.

Stanley, C. (2014). Comparing vector-based and act-r memory models using large-scale datasets: User-customized hashtag and tag prediction on twitter and stackoverflow. *Unpublished doctoral dissertation, Rice University, Houston, TX*, .

Stanley, C., & Byrne, M. D. (2016). Comparing vector-based and bayesian memory models using large-scale datasets: User-generated hashtag and tag prediction on twitter and stack overflow. *Psychological Methods*, *21*, 542–546.

Stevens, J. R., Marewski, J. N., Schooler, L. J., & Gilby, I. C. (2016). Reflections of the social environment in chimpanzee memory: Applying rational analysis beyond humans. *Royal Society Open Science*, *3*, 160293.

Todd, P. M., & Gigerenzer, G. E. (2012). *Ecological rationality: Intelligence in the world.*. Oxford University Press.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive Science*, *42*, 644–691.

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*, *9*, 418–455.

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

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annuual Review of Psychology*, *55*, 235–269.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.

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

Appendix A. Details of Spacing Models

The main text presented only the central predictive equations for each mathematical model. The three spacing models involved components that have their own defining equations:

Appendix A.1. `P&A` *(Pavlik and Anderson)*

Each occurrence of an item has its own decay d_i. The decay for the first presentation was fixed:

$$d_1 = a \tag{3a}$$

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

$$d_n = a + c \times \sum_{j=1}^{n-1} t_j^{-d_j} \tag{3b}$$

Appendix A.2. PPE (Predictive Performance Equation)

PPE requires calculation of elapsed time T_i and the decay rate d_i. The elapsed time is a weighted average of the times t_j that an item has occurred:

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

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, lag_j, between successive presentations determine the value of the decay d_i:

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

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

Appendix A.3. MCM (Multiscale Context Model)

The traces decay according to the rule

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

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:

$$\Delta x_i = 1 - \frac{\sum_{k=1}^{i} \gamma_k x_k}{\sum_{k=1}^{i} \gamma_k} \tag{5b}$$

The original MCM model had a different multiplicative scaling of Δx depending on whether an item is recalled or not. In the case of strings occurring in texts there is not a recall step. Therefore, we assumed a simpler update rule. This has the added advantage of making the model computationally tractable for high frequency cases.

The parameterization of the model concerns the determination of the decay rates τ_i and the relative weights of the traces γ_i. They propose that the decay parameters are monotonically 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:

$$\tau_i = \mu \times \nu^i \tag{5c}$$

$$\gamma_i = \frac{\omega \times \xi^i}{\sum_{j=1}^{N} \xi^i} \tag{5d}$$

With the constraints that $\nu > 1$ and $\omega < 1$ to produce the desired ordering.

Appendix B. Fitting Environmental Data

The data were fit to the log probabilities as displayed in the ”Recency and Frequency” and the ”Once and Twice Occurring” figures (parts a of Figures 5-8).). For each model, probabilities were predicted for a 1000x1000x225 array where the dimensions were (1) delay since the most recent presentation, (2) delay since the second most recent present (if there was more than one presentation), and (3) number of presentations. There were many empty cells in this array because of impossible combinations. The 1000x1000x225 array of probability predictions was then collapsed into probabilities for the points fitted in the figures. For any specific point we calculated a weighted average of all the specific cells in the 1000x1000x225 array that went into that point. The weights were the frequencies of those cells in the data divided by the total frequency of the cells the being collapsed. This weighted-average probability was then converted into a log probability.

In the case of the mathematical models, the predictions for the 1000x1000x225 array were calculated by following the equations of the model. For the GPE and ACT-R models these cells in the 1000x1000x225 array represented the full set of possible predictions. Other models made predictions for a cell that could differ a little depending on the times of presentations that were earlier than the last two. The predictions for these models were based on prototype cases where all of the earlier presentations were evenly spaced. Note that this approximation is irrelevant to the predictions for the "Once and Twice Occurring" data because there are no occurrences before the last two.

In the case of the Monte Carlo models, the generated probabilities were obtained by simulating many items and then averaging all the predicted probabilitiies for the cases that fit in the different cells of the 1000x1000x225 array. Statistics reported in Table 1 are based on the conditions for which there was greater than 5000 observations in the data and greater than 1000 cases in the Monte Carlo simulation.

Appendix C. Evidence for Assumptions in the `A&M` Model with Power Decay

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 7 and 8, there is evidence in the environment specific to each of these assumptions:

a. Distributions of Item Desirability and Decay Rate: Burrell (Burrell & Cane, 1982) chose a gamma distribution for the desirabilities 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 C.1a 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.

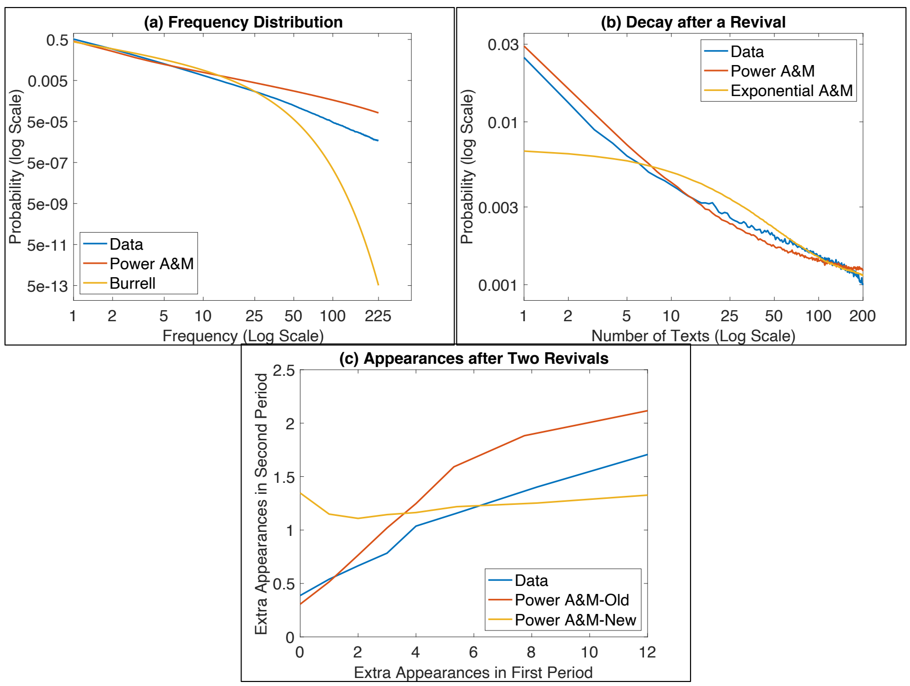

Figure Appendix C.1: Patterns in the environment that reflect the distinct assumptions of the `A&M` power model: (a) Frequency of strings in 1000-text window, compared to the power `A&M` model and Burrell's original model. (b) Probability of a string occurring at various delays after an imputed revival, compared to `A&M` models with power decay and with exponential decay. (c) Number of appearances after a second imputed revival as a function of number of appearances after the first revival, compared to power `A&M` models with old versus new desirability and decay after a revival.

b. Power Law Decay. The linear relationship for frequency in Figure C.1a 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 systems (Mandelbrot, 1959; Simon, 1955). However, Zipf's law does not predict the 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 from 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 investigate here whether there is direct evidence of 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[12]. Then we calculated the 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 C.1b 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. Using sources that allowed 2000-text windows, we identified 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[13]. We counted the number of further mentions in the 200 texts after these first mention in the two windows. Figure C.1c 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. We also obtained predictions from a model identical

[12] 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.

[13] 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.

to the A&M model except that the item got a new random desirability and decay upon revival. This is plotted as Power A&M-New. 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). The figure shows that this relationship does not hold if the revivals have new desirabilities and decays. The function for the Power A&M model 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.

Appendix D. 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 manner (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

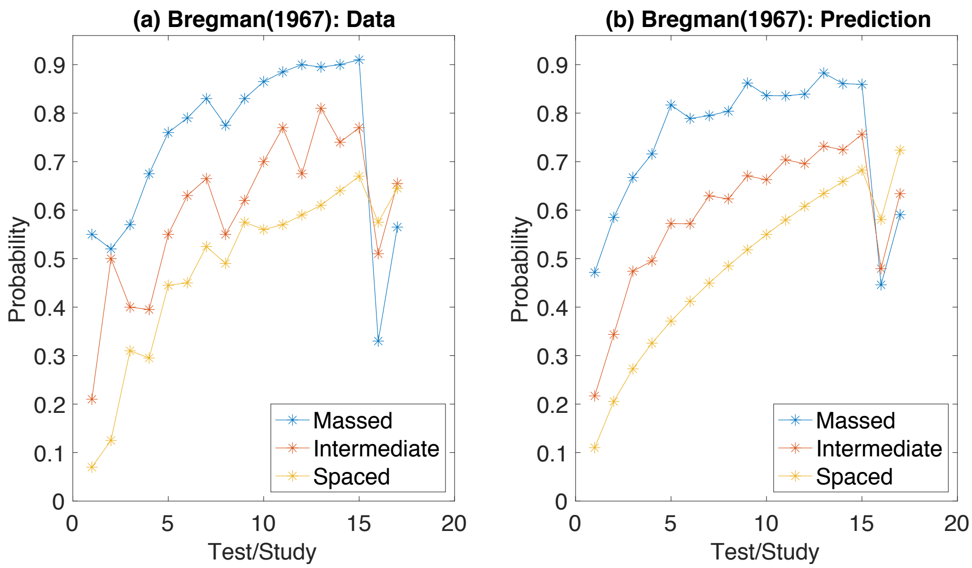

Figure Appendix D.1: Results (a) and AMPE predictions (b) for Bregman (1967). Items receive 15 test-study trials at three levels of test and then are tested at a long delay (last two points).

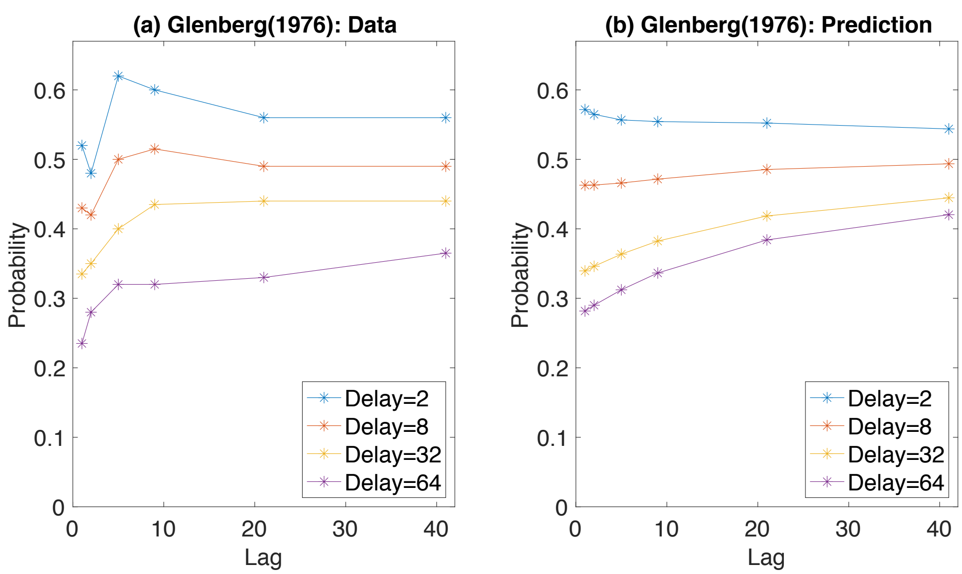

Figure Appendix D.2: Results (a) and AMPE predictions (b) for Glenberg (1976): Effect of different lags between two studies on recall at various lags.

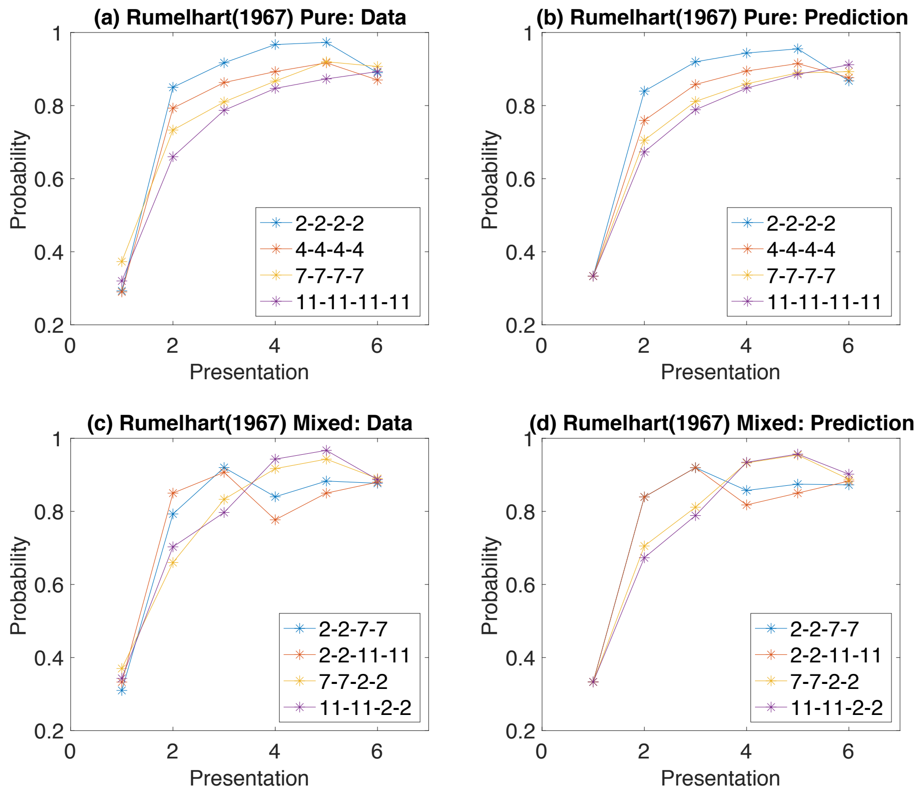

Figure Appendix D.3: Results (a) and AMPE predictions (b) for Rumelhart (1967): Growth in recognition accuracy for items studied at various lags.

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

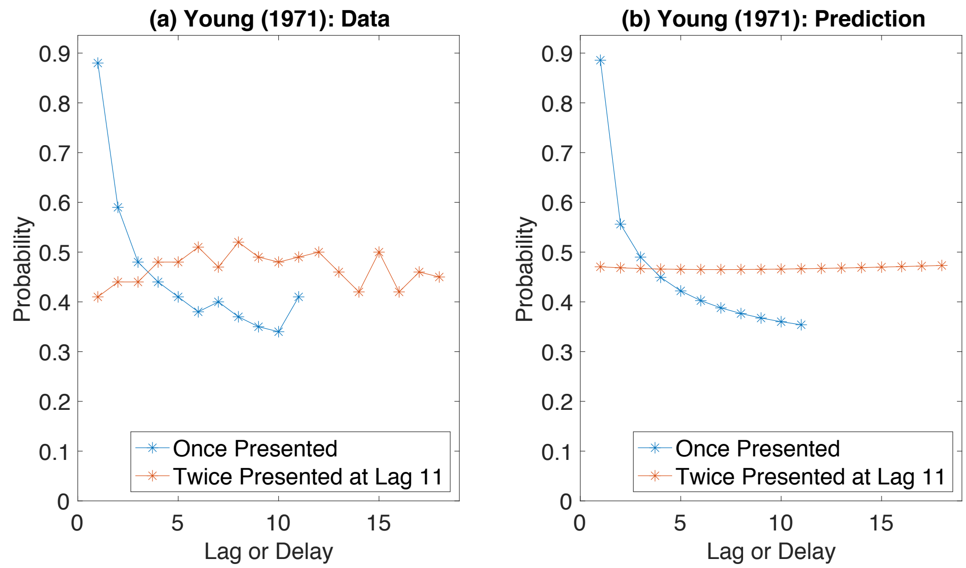

Figure Appendix D.4: Results (a) and AMPE predictions (b) for Young (1971): Recall of items presented once or twice at various lags.

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 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

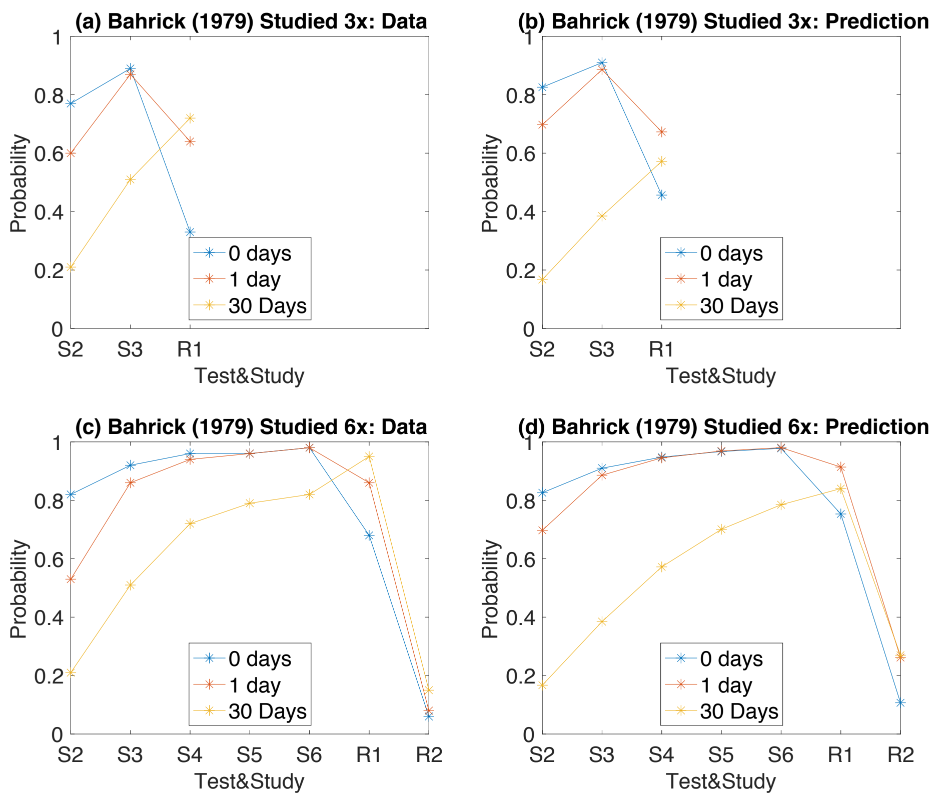

Figure Appendix D.5: Results (a and c) and AMPE predictions (b and d) for Bahrick (1979) and Bahrick & Phelphs (1987). Parts (a) and (b) are for items studied three times at various lags and then tested at a lag of 30. Parts (c) and (d) are for items studied 6 times at various lags and then tested at a lag of 30 and again at a long of 8 years.

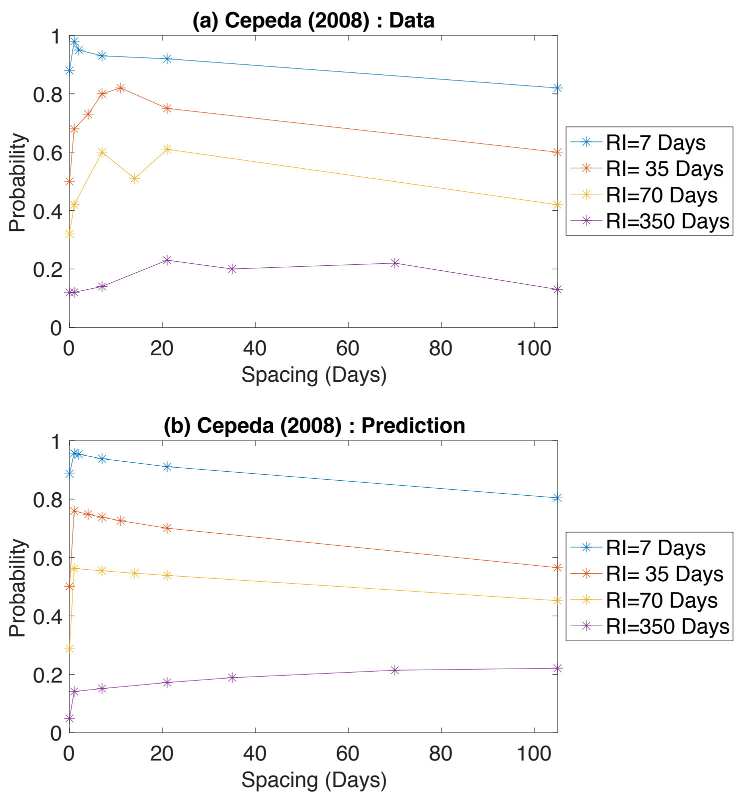

Figure Appendix D.6: Results (a) and AMPE predictions (b) for Cepeda et al. (2008): Retention results at various delays for items studied twice at various lags.

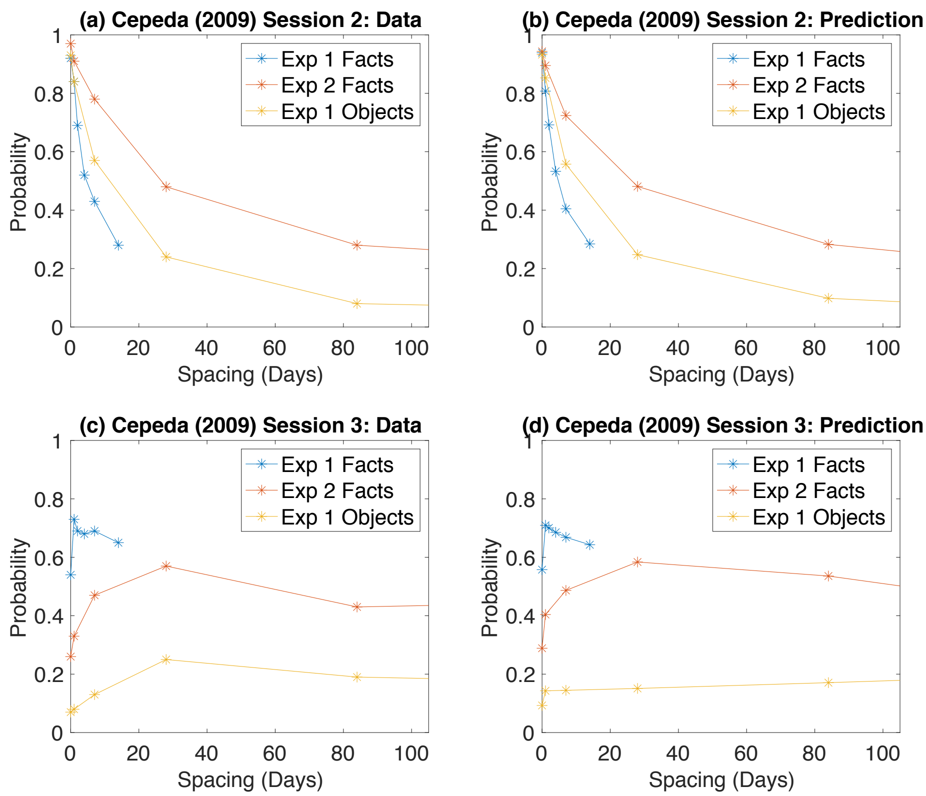

Figure Appendix D.7: Results (a and c) and AMPE predictions (b and d) for Cepeda et al. (2009): Parts (a) and (b) are for the second study session and parts (c) and (d) are for the third retention test.

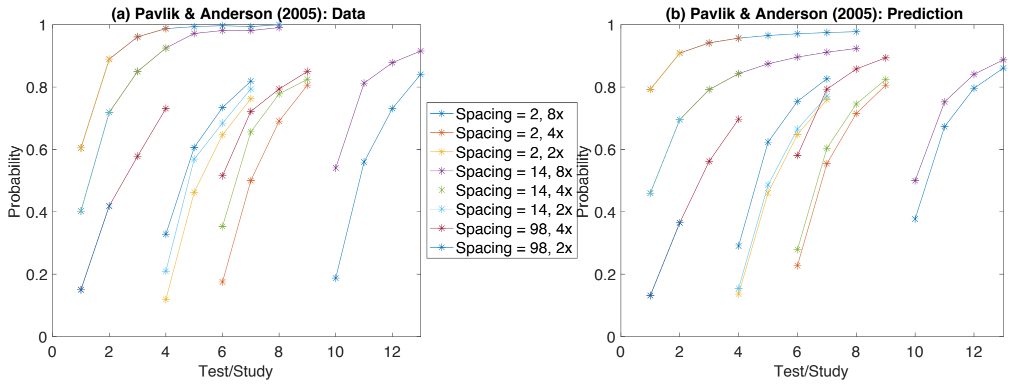

Figure Appendix D.8: Results (a) and AMPE predictions (b) for Pavlik & Anderson (2005): Initial learning of items in the first session at point starts at x=1. The relearning of items in the second session is plotted starting at x=4 if there were 2 tests in session 1, at x=6 if there were 4 tests in session 1, and at x=10 if there were 8 tests.

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 numbers of times at varying lags on one day and then had then perform 4 further test-and-study trials at lags of 98 intervening items either 1 or 7 days later. Altogether there were 162 different conditions crossing different numbers of test-studies (1, 2, 4, or 8) on day 1, lags (2, 14, or 98) between these studies, and 1 vs. 7 days retention. However, Walsh et al. (2018) collapse over days. They also collapsed various observations on day 1 (e.g. performance on the second test-study when there would be 2, 4, or 8 test-studies). We fit AMPE to the same 64 collapsed data points that Walsh et al used (though we did apply the model to the 2 retention intervals and then averaged the results). Compare with Figure 7 in Walsh et al. (2018).

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).

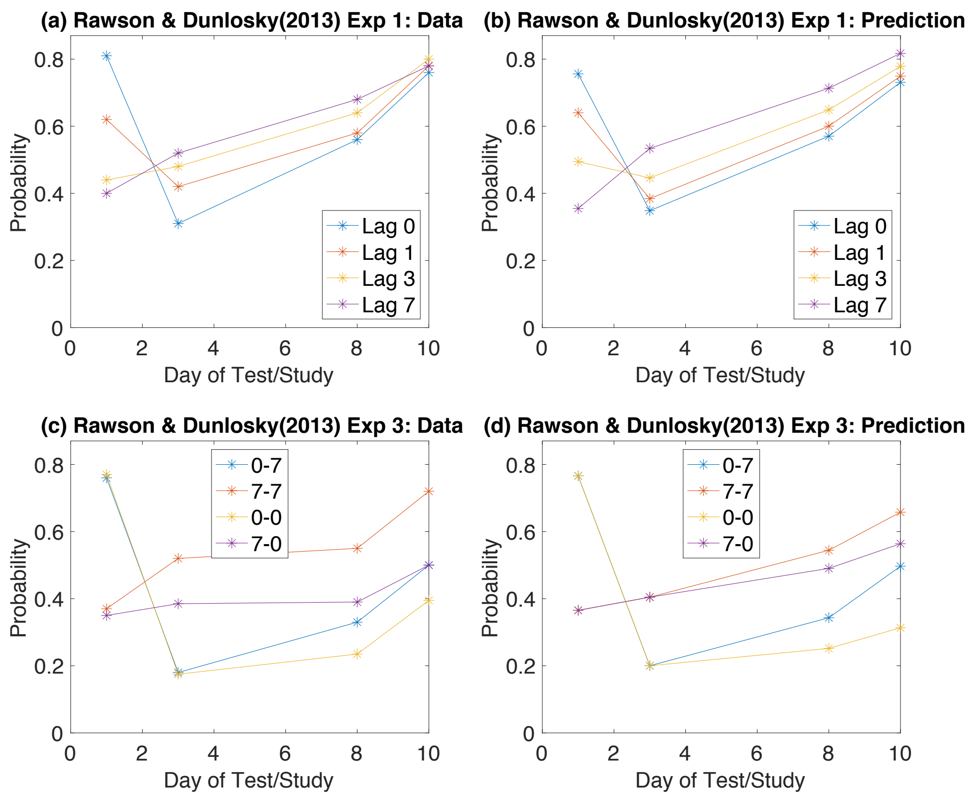

Figure Appendix D.9: Results (a and c) and AMPE predictions (b and d) for citerawson2013relearning. Parts (a) and (b) are for the first experiment that manipulated study lag in the first session. Parts (c) and (d) are for the third experiment that crossed study lag in the first session with study lag in later sessions.