Research Article

REFLECTIONS OF THE ENVIRONMENT IN MEMORY

\17\7

John R. Anderson and Lael J. Schooler
Department of Psychology, Carnegie Mellon University

Abstract—*Availability of human memories for specific items shows reliable relationships to frequency, recency, and pattern of prior exposures to the item. These relationships have defied a systematic theoretical treatment. A number of environmental sources (New York Times, parental speech, electronic mail) are examined to show that the probability that a memory will be needed also shows reliable relationships to frequency, recency, and pattern of prior exposures. Moreover, the environmental relationships are the same as the memory relationships. It is argued that human memory has the form it does because it is adapted to these environmental relationships. Models for both the environment and human memory are described. Among the memory phenomena addressed are the practice function, the retention function, the effect of spacing of practice, and the relationship between degree of practice and retention.*

The title of our paper is inspired by the following remark in Shepard (1990): "We may look into that window on the mind as through a glass darkly, but what we are beginning to discern there looks very much like a reflection of the world" (p. 213). He was commenting on how the principles of perception are exquisitely tuned to the features of the environment in which we live. Basically, Shepard's thesis is that perception has been optimized through evolution to make the best possible inferences about the world given the perceptual input. Recently, Anderson (1989, 1990) has suggested that the same might be true about human memory.

Many people hold the bias that human memory is anything but optimal. They point to the many frustrating failures of memory. However, these criticisms fail to appreciate the task before human memory, which is to try to manage a huge stockpile of memories. In any system responsible for managing a vast data base there must be failures of retrieval. It is just too expensive to maintain access to an unbounded number of items.

Given the initial bias against human memory, it would be particularly compelling if we could show that human memory were optimal. How does a system behave optimally when it is faced with a huge data base of items and cannot make all of them instantaneously available? It would be behaving optimally if it made most available those items that were most likely to be needed.

In this paper we explore the issue of whether human memory is behaving optimally with respect to the pattern of past information presentation. Each item in memory has had some history of past use. For instance, our memory for one person's name may not have been used in the past month but might have been used five times in the month previous to that. What is the probability that the memory will be needed (used) during the current day? Memory would be behaving optimally if it made this memory less available than memories that were more likely to be used but made it more available than less likely memories.

In this paper we examine a number of environmental sources to determine how probability of a memory being needed varies with pattern of past use. However, we first review how availability in human memory varies with pattern of past use. Some aspects of this problem have been extensively studied in empirical studies of human memory.

FORM OF THE MEMORY FUNCTIONS

Two of the most basic statistics we might gather about pattern of past use are how often a memory has been practiced and how long it has been since it was last practiced. Learning functions and retention functions to describe these two aspects of human memory have been collected since the original experiments of Ebbinghaus (1885/1964). Figure 1 shows the retention function and practice function obtained by Ebbinghaus.

The Retention Function

Ebbinghaus measured retention in terms of the percent savings in relearning a list of nonsense syllables. The function shows the classic negative acceleration typical of such retention functions. In order to be able to compare this memory function to the environment, we need to decide how to characterize the forgetting function. Some (e.g., Loftus, 1985) have suggested that these functions satisfy an exponential formula:

$$P = Ae^{-bT} \qquad (1)$$

where P is the performance measure, T is the delay time, and A and b are parameters of the model. The intuitive appeal of an exponential function probably explains why it is so often suggested. It implies that during each unit of time, the memory loses a constant fraction of what is left. This process evokes images of radioactive decay, an analogy often used to describe forgetting. One can investigate whether this function holds by performing a log transformation of the performance scale. If the underlying relationship is exponential, a linear relationship should obtain between log performance and time:

$$\log P = \log A - bT. \qquad (2)$$

A precondition to performing an adequate test of such a function is that we have a large manipulation of the time scale.

Tuesday, February 11, 2003 (3).max
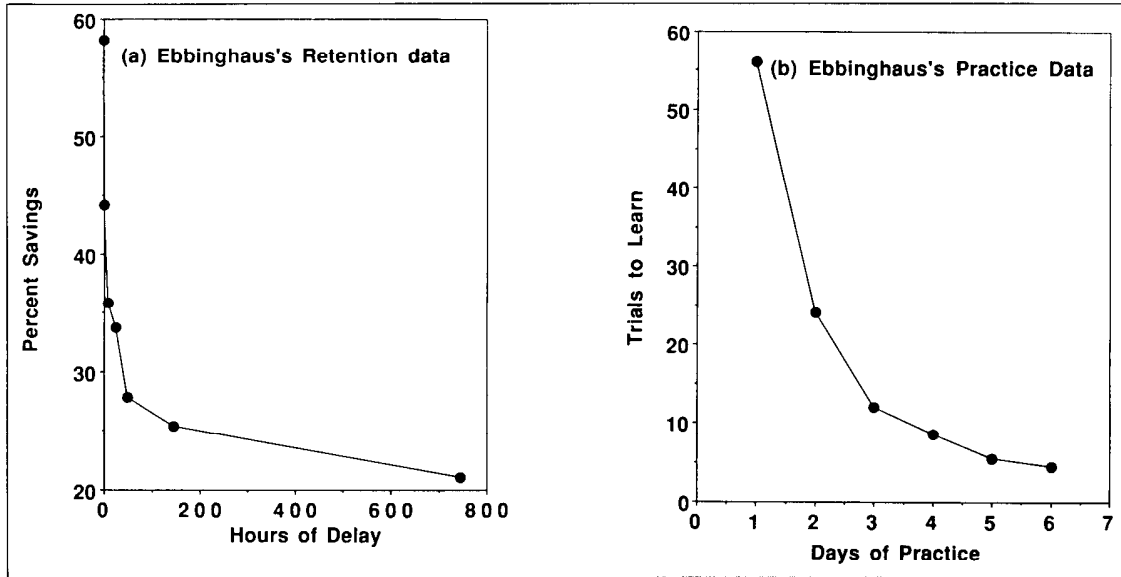
John R. Anderson and Lael J. Schooler

Fig. 1. (a) Ebbinghaus's (1885/1964) retention function showing percent savings as a function of delay. Ebbinghaus used delays from 20 minutes to 31 days. (b) Ebbinghaus's practice data showing total number of trails to master a set of lists as a function of number of days of practice.

Ebbinghaus's data certainly satisfy this precondition, as he varied retention intervals from 20 minutes to 31 days.

Figure 2a illustrates the Ebbinghaus data with the performance scale transformed. As may be observed, the resulting function is anything but linear. Thus, despite its popularity, the hypothesis of an exponential forgetting function is not supported. Wickelgren (1976), using a d' memory measure and delays from 2 minutes to 14 days, found evidence for a power function relating delay to retention.[1] A power function has the form:

$$P = AT^{-b}. \tag{3}$$

_____

1. Actually, Wickelgren's theory also had an exponential component that would dominate the power component at very long delays.
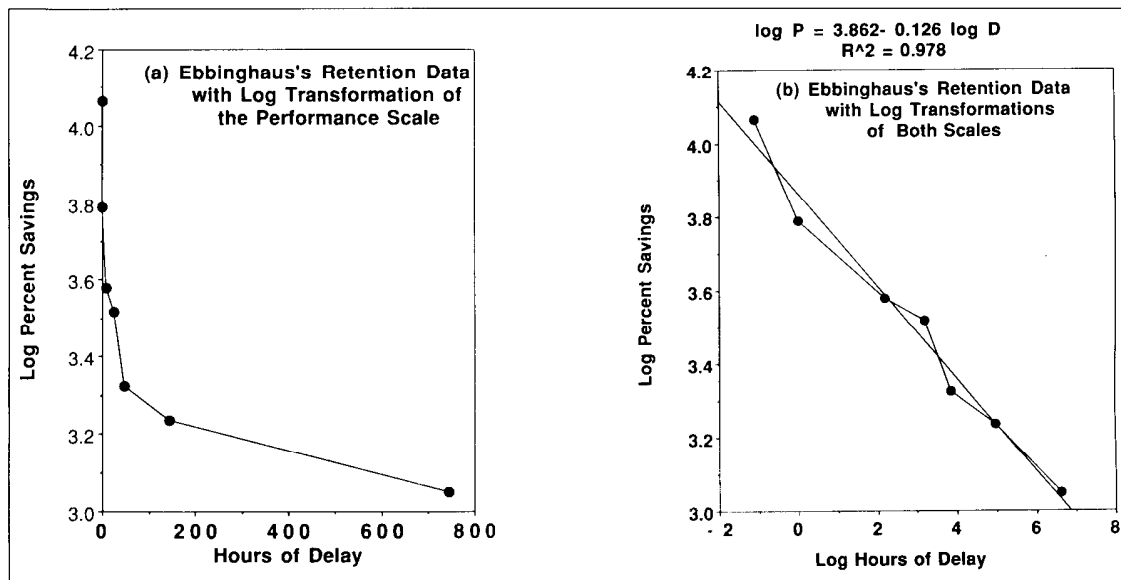

Fig. 2. The retention data from Figure 1 with (a) the performance measure transformed according to a logarithmic function and (b) both performance and delay scales transformed according to a logarithmic function.

Reflections of Environment in Memory

This can produce a very slowly decaying memory function. If one performs log transformations of both the performance measure and the time measure, one obtains a linear relationship:

$$\text{Log } P = \log A - b \log T. \tag{4}$$

Figure 2b illustrates the Ebbinghaus data with both scales log transformed. As can be seen, one gets a very good approximation to a linear relationship in these log scales with log A = 3.862 and b = −.126. If we go back to the original scales, we get a relationship of the form:

$$P = 47.56 \, T^{-.126}. \tag{5}$$

The exponent .126 can be taken as the forgetting rate.

A power function implies that the performance measure will go to infinity as time goes to zero. In contrast, an exponential function implies a bound on how good performance can be at t = 0. Although we never realize a true delay of zero, we still can fail to find power functions if we use scales with an upper bound. Probability of recall is such a scale. Ebbinghaus's percent savings is another scale, but even at the 20-minute delay in Ebbinghaus's experiment there was only 58% savings, so the ceiling was not approached. Power functions for forgetting tend to be obtained when we use measures that do not have upper bounds or do not approach their upper bounds. The d′ measure of Wickelgren is a scale that does not have an artificial upper bounds. Later we will also advocate recall odds rather than recall probability, since odds varies from zero to infinity. Recall time is another measure that ranges from zero to infinity and tends to yield power functions for retention, although in this case we have to switch the sign of the exponent since recall time increases with delay.

One of our goals is to explain why retention functions tend to satisfy a power relationship. Given that people have preferred an exponential function on an intuitive basis, such an explanation would be a nontrivial result. Power functions seem to describe memory performance from a few seconds to years. As Wickelgren (1974) has argued, there does not seem to be any discontinuity that would be associated with a shift from short-term memory to long-term memory. It will be a significant result if we can find a reason for predicting a power function (in contrast to an exponential function) from an analysis of the environment.

The Practice Function

We can ask the same thing about the practice functions—are they better fit by an exponential form or a power form? The measure used in Ebbinghaus's Figure 1b is appropriate for addressing this question. Plotted there are the number of trials to learn a list of 36 nonsense syllables to a criterion of one correct anticipation. Ebbinghaus practiced these lists each successive day and so we see the improvement across days with practice.

Figure 3 compares how well exponential and power functions fit these data. The range of practice (1 to 6 days) is not large enough to enable a clear discrimination among the functions, although the power function produces a somewhat better fit. This practice function has been explored over much larger ranges of practice and a power function typically provides a better fit (Newell & Rosenbloom, 1981), although there has
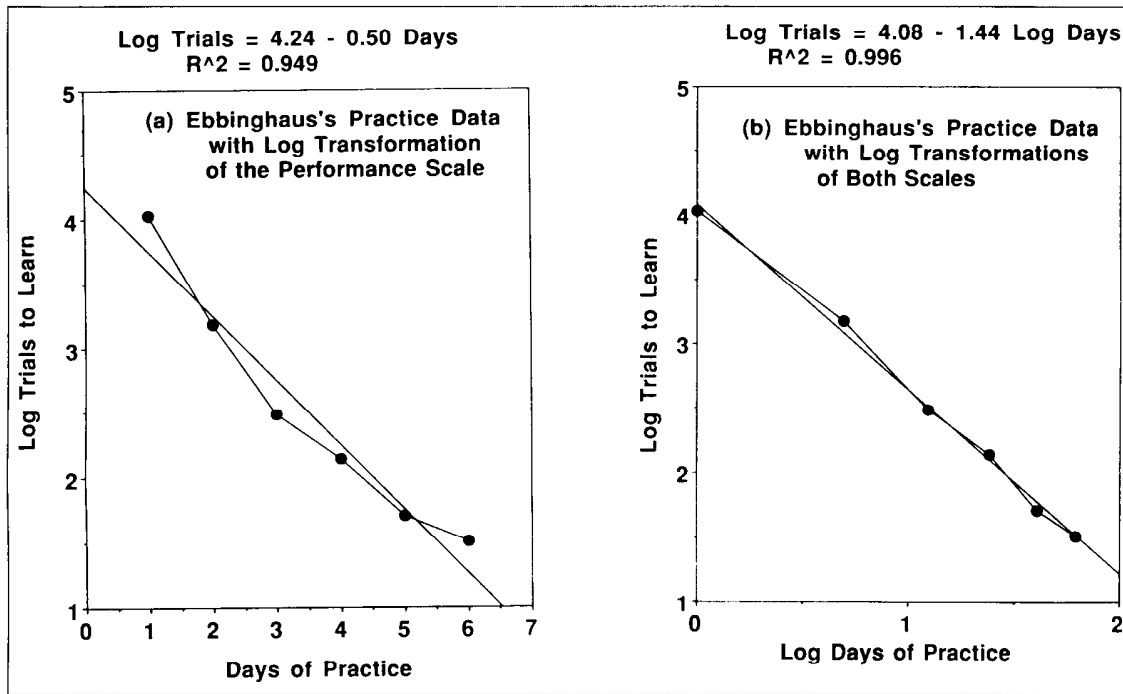
Fig. 3. The practice data from Figure 1b with (a) the performance measure transformed according to logarithmic function and (b) both performance measures transformed according to a logarithmic function.

again been a history of initial preference for the exponential function (Mazur & Hastie, 1975; Restle & Greeno, 1970). Another goal we have is to provide an environmental explanation for why there is this ubiquitous practice function. Again this result is not trivial given the initial beliefs that the learning function should be exponential in form.

The power function that corresponds to the data in Figure 3 is:

$$P = 513 \, S^{-1.24} \qquad (6)$$

where S is the number of days of study. The size of this exponent can be interpreted as the learning rate.

Implications of Power Functions

Note that in Figures 1–3 we are measuring retention by a savings measure, where larger numbers are better, while we are measuring practice by a trials-to-relearn measure, where large numbers are worse. Throughout the literature one can find a variety of performance scales, some of which have a positive valence like savings and others of which have a negative valence like trials to relearn. Later we have more to say about percent correct, the most common positive valence scale, and reaction time, the most common negative valence scale. Generally, power functions are found whatever scale is used (provided it is not a scale with an upper bound, or if it is, the upper bound is not approached). Forgetting functions display a negative slope on positive valence scales and a positive slope on negative valence scales. This relation is reversed for practice functions. It might seem curious that power functions appear for different performance scales, but the power relationship is a strong one and will be approximately maintained by many transformations of scale. As a final comment, we should say we have no investment in the claim that these empirical functions are best modeled or correctly modeled as power functions. For our purposes, it is enough to note that power functions give remarkably good approximations. Our goal is to show that these remarkably good approximations are implied by the structure of the environmental input to memory.

A number of recent theories are capable of accounting for power-law learning (Anderson, 1982; Lewis, 1978; Logan, 1988; McKay, 1988; Newell & Rosenbloom, 1981; Shrager, Hogg, & Huberman, 1988). Except for Anderson (1982), however, none of these theories are capable of accounting for the forgetting function. This is a serious deficit. Any extended practice must be taking place over many days and it is reasonable to assume that subjects are forgetting the impact of the early training. Models that predict power-law learning but ignore forgetting might well fail to predict power-law learning when forgetting is factored in. The model in Anderson (1982) basically assumes that the power-law learning function arises from a simple linear learning process being slowed down by forgetting. That theory led to the prediction that the forgetting exponent and the practice exponent should sum to 1 (see Anderson, 1982, for derivation). There is no evidence for that prediction in the Ebbinghaus (1885/1964) data nor in any other experimental effort that has obtained estimates of both practice effects and forgetting effects.

Thus, it is fair to say that there is no theory of human memory that adequately predicts both the practice and forgetting functions. This is a pretty startling result since it has been a field of constant research and theorizing for over 100 years.

The Spacing Effect

One other effect that we would like to note creates even greater stress on theories of memory—the spacing effect (Bahrick, 1979; Glenberg, 1976). It is found that the spacing between successive repetitions of an item affects how well the item is remembered. Moreover, this effect interacts with the delay between the last study of an item and the test. Figure 4 displays the results from Glenberg (1976). In this experiment there were two studies of an item followed by a test. The data are organized according to the lag between the two studies and the lag between the second study and the test. At short test lags, recall is better the shorter the study lag. This can be seen as derivative from what we have seen about the retention curve. The longer the study lag, the greater the retention interval from the first study to the final test. However, when the test lag is long, there is better recall the longer the study lag. This result contradicts what we would extrapolate from the retention curve alone. The spacing effects might be characterized as showing greatest recall when study lag matches test lag. Whether this conclusion is correct or not is unclear, but there is abundant evidence for an interaction of the sort illustrated in Figure 4 between study lag and test lag.

No theory of human memory, including Anderson (1982), has been able to account adequately for practice effects, retention effects, and the spacing effect. The reason should be ap-
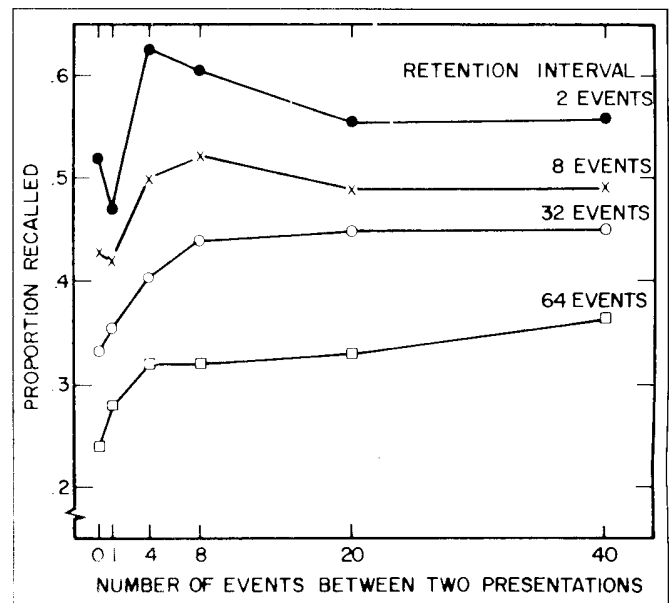

Fig. 4. The proportion of paired-associate responses recalled as a function of the number of events between two presentations of repeated items (lag interval) and the number of events between the second presentation and the test (retention interval). From Glenberg (1976).

parent, as the three effects would seem to be somewhat in contradiction. Holding test lag to last presentation constant, the advantage of each presentation should diminish as they are spaced further apart because we are increasing the retention intervals from the earlier presentations. However, the spacing effect tells us that this is not always true. One should not think, however, that the spacing effect eliminates the retention effect. The biggest effect in Figure 4 is the retention effect, which is reflected in how far apart the separate curves are. One way of characterizing what is going on is that there is a large effect of delay since last presentation but that the other delays have a less clear effect.

A number of theories have been able to predict simultaneously a forgetting function, a retention function, and a spacing effect (Estes, 1955; Landauer, 1975; Glenberg, 1976). However, it does not seem that they can predict the power-function form that these functions appear to take. These theories assume that memories get associated to contexts that gradually change over time. The practice function simply results from the increased associations to context with repetition. The retention function results because with time, the test context changes from the learning context. The spacing effects result because at long lags memories are likely to be associated to different contexts. This results in increased probability that the test context will overlap with one of the study contexts. Such a model might well be given an expression that would produce the parametric form of the three effects. However, we and others have been frustrated in our attempts to find such an expression.[2]

AN ENVIRONMENTAL EXPLANATION

Given that there have been no successful mechanistic explanations for practice, retention, and spacing phenomena, it becomes all the more interesting to see whether we can explain these phenomena from the assumption that the memory system is adapted to the structure of the environment. The basic idea is that at any point in time, memories vary in how likely they are to be needed and the memory system tries to make available those memories that are most likely to be useful. The memory system can use the past history of use of a memory to estimate whether the memory is likely to be needed now. This view sees human memory in some sense as making a statistical inference. However, it does not imply that memory is explicitly engaged in statistical computations. Rather, the claim is that whatever memory is doing parallels a correct statistical inference.

What memory is inferring is something we call the *need probability*, which is the probability that we will need a particular memory trace now. The basic assumption developed in Anderson (1990) is that memories are considered in order of their need probabilities until the need probability is so low that it no longer is worth considering any more. If we let p be the need probability, C be the cost of considering a memory, and G

---

2. Wickelgren (1972) produced a mathematical theory that was tailored to the form of the retention function but does not address the form of practice function. It mispredicts the spacing effect in that it claims that the utility of later presentations is a function of how distant they are from the first. It has no role for the lag among these later presentations.

be the gain associated with a successful retrieval, one should stop when C > pG.

Despite the description of this process in terms that evoke images of memories being considered one at a time, there are equivalent parallel processes. We prefer a parallel model in which different memories are allocated different resources according to their need probability. However, for current purposes we simply note that this analysis does not imply a commitment as to the mechanism of retrieval.

Relationship between Need Odds and Behavioral Measures

This analysis does allow predictions to be derived about the relationship between need probability and the dependent measures of recall latency and recall accuracy. With respect to recall latency, the critical assumption is that there is a distribution of memories in terms of their estimated need probabilities. The reasonable assumption is that there will be a mass of need probabilities near zero with a tail of a few higher probability memories; that is, to say the distribution of memories will be J-shaped or highly skewed. It is more convenient to think about the shape of such a distribution in terms of *need odds*. If p is need probability, then $q = p/(1 - p)$ will be need odds. An odds measure has the advantage of varying from zero to infinity. Thus, the expectation is that most memories will have near-zero odds and a rapidly diminishing few will have higher odds.

A great many phenomena show such J-shaped distributions, including distributions of scientists by number of publications, words by frequency, and firms by size. Simon and Ijiri (1977) present the following density as characterizing such distributions:

$$f(x) = ax^{-k} \tag{7}$$

where f is the frequency of an item of measure x (e.g., word frequency, firm size, or need odds) and a and k are constants.

If we assume that memories are examined in order of odds, then the time to examine a memory with odds q will be proportional to the number of memories with odds greater than q. This can be calculated as:

$$\int_q ax^{-k}dx = bq^{-(k-1)} \tag{8}$$

where $b = a/(k - 1)$. Thus, we see that time is related to need odds as a power function with exponent $(k - 1)$. Thus, if odds were related to retention interval or practice as a power relation with exponent c, then time would be related to retention interval or practice with exponent $c(k - 1)$. The force of this analysis is that power functions in need probability imply power functions in time, although not necessarily with the same exponent. If $k = 2$, the exponent will be the same. Simon and Ijiri report that values of $k = 2$ are common.

The above was an analysis of time. Anderson and Milson (1989) can be consulted for a similar analysis of recall probability. The basic assumption there is that recall will stop before retrieving the target item if its need probability is too low. This

John R. Anderson and Lael J. Schooler

might seem to imply a step function in which all items above a certain need probability are recalled and all below are not recalled. However, there has to be some noise in the process such that the distance between an item's need probability and the threshold varies. A natural scale on which to try to model this variation is log need odds, which varies from minus infinity to infinity. If we assume that there is a normal distribution of estimated log need odds around true need odds, we predict a sigmoidal function rather than a step function relating need odds to recall odds. Anderson and Milson show that this relation implies a power relationship between need odds and recall odds. Thus, as in the case of time, we see that the natural prediction is that a power function in need odds implies a power function in the observed behavior. Again, the exponent need not be the same.

These considerations about recall odds and reaction time greatly simplify our research program. They mean that these dependent measures should directly reflect the functional form and ordinal relationships displayed by need odds. Thus, we can look to see whether need odds functions are power functions like the behavioral functions. It is not necessary that they have the same parameters such as exponent or scale constant for the power function. For instance, it is reasonable to suppose recall odds will be much greater than the corresponding need odds, but they should have the same functional forms.

INFORMATION ABOUT ENVIRONMENTAL STRUCTURE

What we need to find out is how past history of usage of information predicts the probability that the knowledge will be used in the next time interval. Anderson and Milson (1989) developed a theory based on mathematical models that were developed to explain library borrowings and accesses to files in computer systems. While this approach has some strengths, it has two considerable weaknesses that we hope to redress in this paper. First, while these are examples of systems that have to retrieve information, they are not systems facing human retrieval demands and so we are left with an argument by analogy. Second, while a formal model has some analytic advantages, it obscures the very direct relationship being proposed between the environment and memory, leading some (e.g., Simon, in press) to claim that the predictions rest on the auxiliary assumptions in the environmental model. Quite the contrary, it is the case that the predictions are a direct reflection of the structure of the environment.

Ideally, we would like to follow people about determining when demands are being made on their memory to retrieve a piece of information and how demands for the same piece of information tend to repeat over time. While this is technically infeasible, it is possible to study certain subsets of demands that are being placed on human memory. We have studied the following three sources (see Schooler and Anderson, unpublished, for detailed information about each source):

1. We have analyzed 730 days of New York Times headlines from January 1, 1986, to December 31, 1987. Every time a particular word like "Qaddafi" appears in the New York Times headline, this is a demand on a potential reader of the article to retrieve information about the referent of that word to decide whether this is an article that the reader might want to read.

2. We have looked at the subset of the CHILDES data base of MacWhinney and Snow (1990) looking at children's verbal interactions. Every time someone says a word to a child, this is a demand on the child to retrieve the word's meaning.

3. We have looked at the electronic mail messages the first author (J.A.) received from March 1985 to December 1989. Here we have analyzed the senders of the messages. The assumption here is that every time J.A. receives a message from a certain person, that is another demand to retrieve some information from J.A.'s memory about the sender.

Figure 5 illustrates the pattern of usage of some words over a 100-day period for the New York Times. The question of interest is how does this pattern of use over the 100 days predict the probability of use on the 101st day? In addressing this question we can look at the relationship between various statistics describing the past 100 days and probability of occurring on the 101st day. For instance, "Reagan" occurs 52 times in that 100-day period. We can look at need probability on the 101st day. This would be representative of an item that has had 52 practices in an experiment and we are looking at its recall. It turns out in this case "Reagan" actually appeared in the headlines on day 101 but aggregating over items that appeared 52 times in a 100-day window, some will appear on day 101 and some will not. We can use the empirical proportion as an estimate of the probability that an item used 52 times in 100 days will be needed on day 101.

The Practice Function

Figure 6a shows the relationship between the number of previous days on which a word has appeared during the past 100
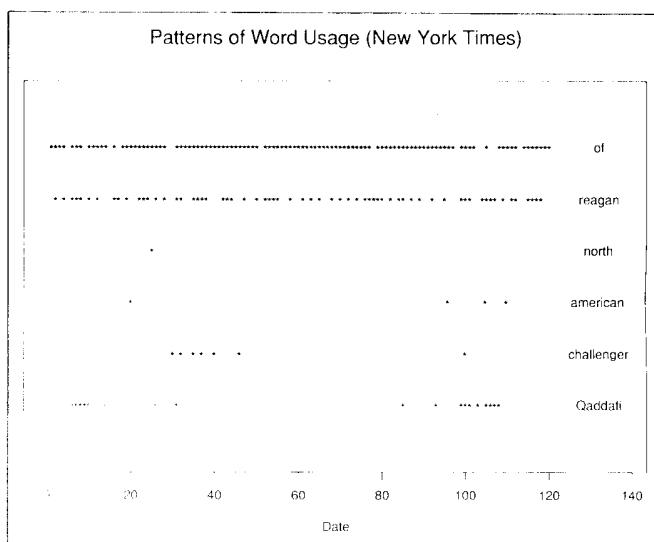

Fig. 5. Patterns of usage of various words in the New York Times data base over a 100-day period.
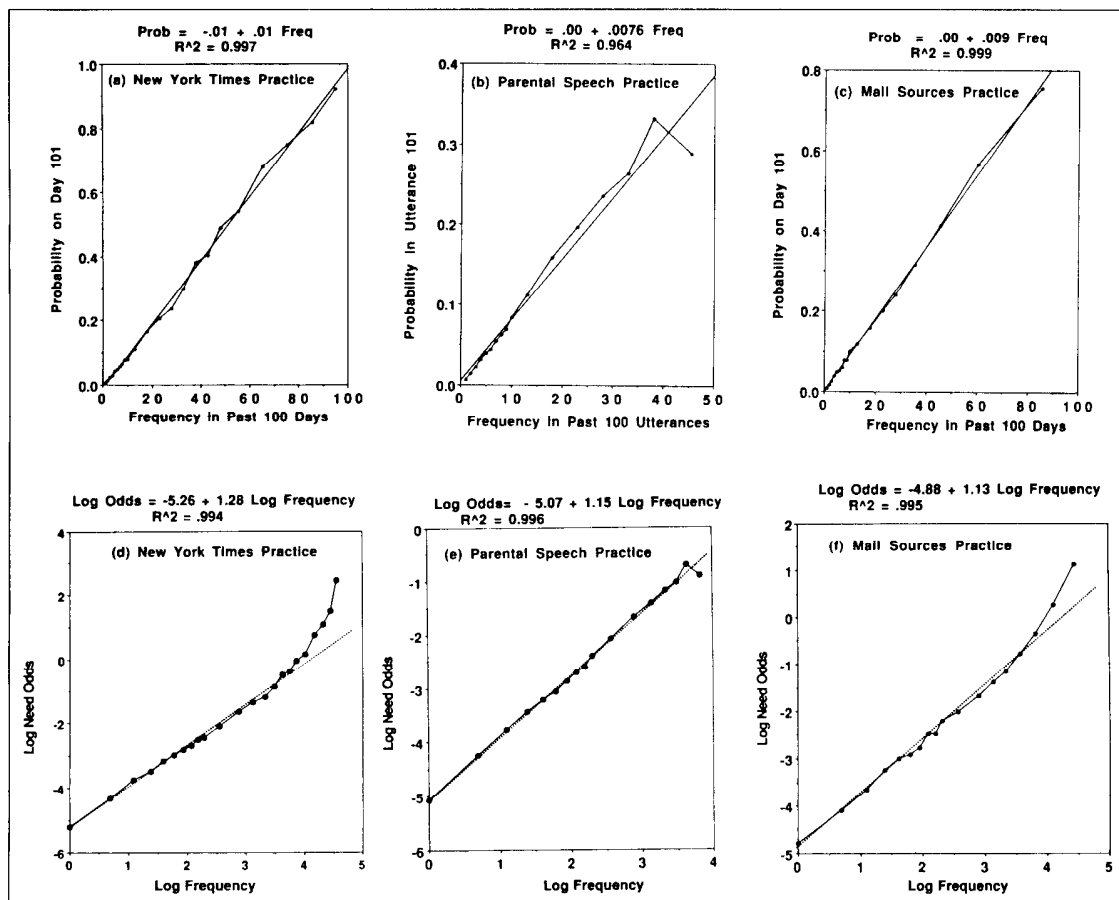
Reflections of Environment in Memory

Fig. 6. (a) Probability of a word occurring in a headline of the *New York Times* on Day 101 as a function of the number of times it occurred in the previous 100 days; (b) probability of a word occurring in the 101st utterance from a parent as a function of the number of times it occurred in the previous 100 days; (c) probability of receiving a message on the 101st day from a source as a function of the number of times messages were received from that source in the previous 100 days. Panels (d–f) provide transformation of (a–c) plotting log needs against log frequency.

days and the probability it will appear in the current day. We have plotted probability of occurrence on the 101st day against number of uses in the previous 100 days. This analysis reveals a particularly straightforward relationship. In this data base, future probability of use perfectly reflects the proportion of past use in the data base.

Figure 6b shows a similar analysis for the child language data base. Here we are looking at the probability of a word occurring in the 101st utterance to the child as a function of the number of times it appeared in the previous 100 utterances to the child. Again we have plotted probability of use against number of prior utterances. The relationship is again linear, although we find that past proportion overestimates future use. Basically, if an item has occurred in a proportion P of the past 100 utterances, it has a probability .76P of occurring in the next utterance.

Finally, Figure 6c shows a similar analysis for the electronic mail data. Again a linear relationship is found, but this time the function is .9P.

Simon (1955) noted that the probability of an item being repeated was proportional to its past frequency of usage in a number of sources. We have just replicated this result. The constant of proportionality (1.0 for *New York Times*, .76 for child language, and .9 for mail messages) reflects the rate at which new terms are appearing. One minus this constant is the probability that the next item is a new term.

In Figures 6a–c we have plotted the relationship between need probability and frequency. Our prediction is that there should be a power relationship between need odds and frequency or a linear relationship between log need odds and log frequency. Figures 6d–f plot log odds rather than log probability. Generally, there is a strong correlation between log need odds and log frequency but systematic deviations appear for frequencies over 50. We have estimated best-fitting linear functions for frequencies under 50 and the results are every bit as good as in the original Figures 6a–c. We are not bothered by deviations for frequencies over 50 because these represent very few items. In the case of the *New York Times*, they are a few

John R. Anderson and Lael J. Schooler

functor words. In the case of electronic mail, they are two individuals. There are no such items in the case of child language. These few items do represent extremes that are not realized in memory experiments that produce power functions. They are items that occur nearly every day of our lives and no memory experiment comes close to creating that ubiquitous a learning experience.

The Retention Function

We also used a window of 100 days in analyzing the *New York Times* for an analog of the retention function. Here we look at probability of recall on the 101st day as a function of how many days have elapsed since the item last occurred in that 100-day window. Figure 7a shows this relationship with an untransformed scale, and Figure 7d shows the relationship plotting log need odds against log time. As can be seen, the data in Figure 7a show the typical negative acceleration of a retention curve, and the data in Figure 7d show that this satisfies a power function with exponent .73. Figures 7b and e show the compa-

rable analysis from the child language data. Here we plotted probability that the word would appear in the 101st utterance to the child as a function of where last it appeared in the last 100 utterances. Figure 7e shows another power relationship, this time with exponent .77. Figures 7c and f show the data for the mail messages. Again a linear relationship appears in the case of the log transformed data in Figure 7f, implying a power relationship. In this case the exponent is .83. Although we have not bothered to include the plots, in each case the data do not satisfy an exponential relationship.

Spacing Effects

We tried to find an analog of the Glenberg study in the environment. For the *New York Times*, we selected cases where a word occurred exactly twice in the past 100 days and considered the probability of its occurring on day 101. We analyzed this probability of occurrence as a function of the lag between the two occurrences (the analog of study lag) and the lag between the second occurrence and test (the analog of test lag).
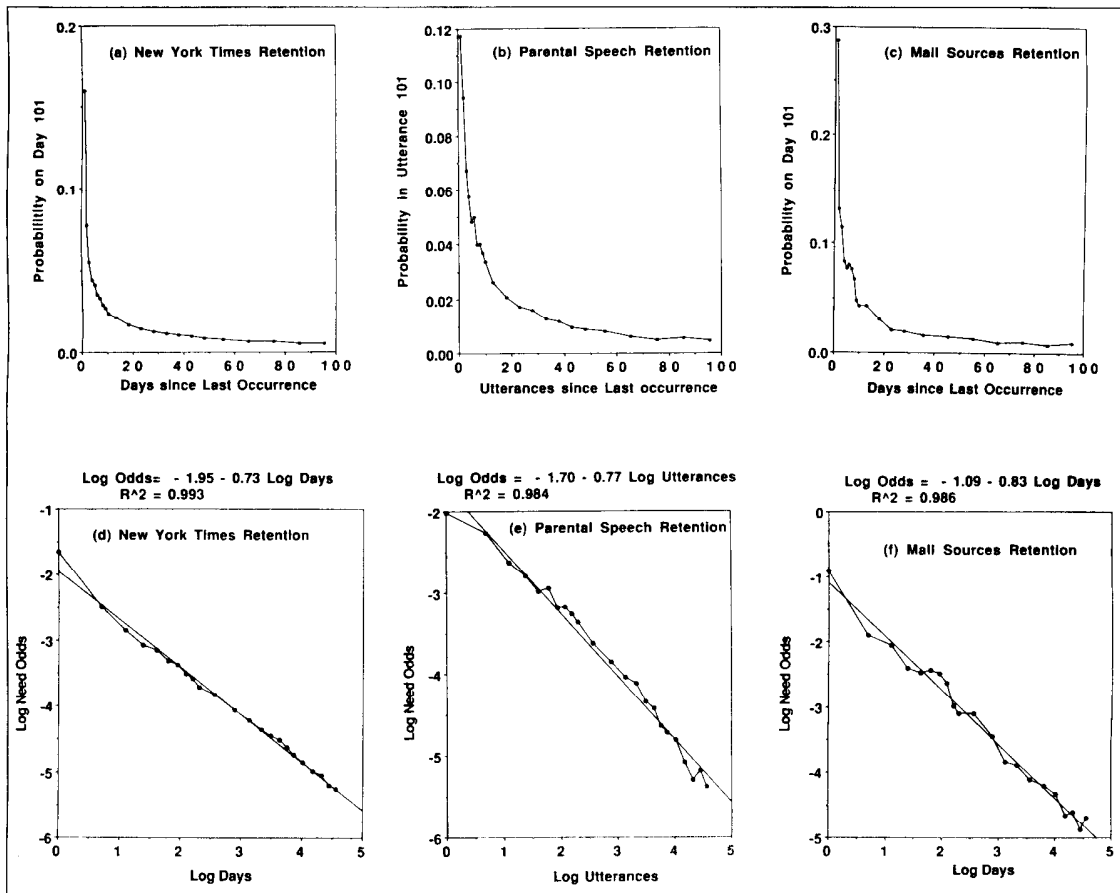

Fig. 7. (a) Probability of a word occurring in a headline in the *New York Times* on day 101 as a function of how long it has been since the word previously occurred; (b) probability of word occurring in the 101st utterance from a parent as a function of how many utterances it has been since the word previously occurred; (c) probability of receiving a mail message from a source as a function of how many days it has been since a message was last received from that source. Panels (d–f) provide transformation of (a–c) plotting log need odds against log frequency.
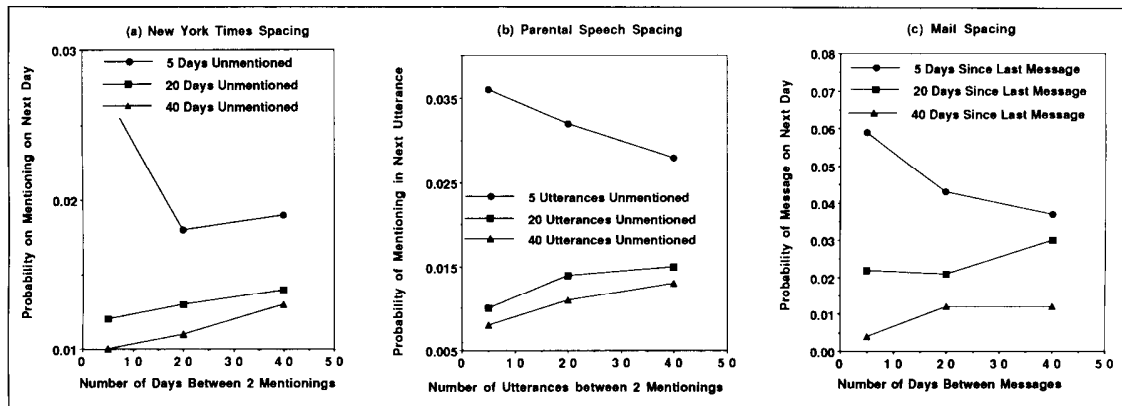
Reflections of Environment in Memory

Fig. 8. Analog of the Glenberg study (Fig. 4) in the (a) *New York Times*, (b) child language data source, and (c) electronic mail data source.

Such data are relatively rare, and therefore we collapsed these into three categories—lags of 1–9, lags of 10–30, and lags of 31–89. Classifying study and test lag according to these three categories gives us a 3 × 3 classification of the data. Figure 8a shows the data organized according to this classification. This figure qualitatively reproduces the data of Glenberg. At short test lags, probability decreases with study lag, but the reverse relationship holds for long test lags. Figure 8b shows the same analysis for the child language data, and Figure 8c shows the data for mail messages. Again the same qualitative interaction appears.[3]

The data in both figures are plotted as Glenberg reports his data—the abscissa is study lag and different curves represent different retention lags. This analysis makes the point that, as in Glenberg's data, the big effect is for the retention interval (different curves) and the relatively small effect is for the study lag (shape of individual curves). It is of interest to replot this data looking for the effect of retention interval for various study lags. We have done this in Figure 9, collapsing the two longer study intervals. Two things are apparent. First, the retention function is steeper for the shorter study lag (.20 for long lag and .49 for short lag in *New York Times*; .45 for long lag and .76 for short lag in parental speech; .48 for long lag and 1.03 for short lag in mail sources). Second, these functions, which are controlled for number of prior studies, show much shallower slopes than those in Figure 7, where it was possible that number of prior studies was confounded with retention interval. The shallower slope is particularly apparent in the case of long lags.

Interactions between Practice and Forgetting

Recently, there has been some controversy as to the form of various forgetting functions at various degrees of learning (Bogartz, 1990; Loftus, 1985; Slamecka & McElree, 1983). Unfor-

tunately, this research has not considered the power function for forgetting.[4] Figure 10 shows some of the data that have fueled this controversy on graphs that plots log odds scales. Figure 10a is from Hellyer (1962), who gave one to eight presentations of a three-consonant unit followed by a retention interval of 3 to 27 seconds. Figure 10b is from Krueger (1929), who trained a list of 12 nouns to various degrees of overlearning and then looked at retention from 1 to 28 days. Figure 10c is from Underwood and Keppel (1963), who looked at retention of nine letter associates at 1 or 7 days as a function of number of trials of training. Figures 10a and b use log delay as the abscissa and plot different degrees of learning as different curves. Figure 10c plots amount of learning as the abscissa and has two different curves for the two different delays. All three sources illustrate the same point. Delay and practice have approximately additive effects in these log transformed scales.

One interesting question is what is the relationship in our three environmental sources. Figure 11a shows the retention data from the *New York Times* broken down into high- and low-frequency items and Figure 11b shows the comparable data for child language. Both data sources show the same approximately additive effect of the two factors. We should stress that we have no investment in the claim that the effects are truly additive in either memory or the environment. Rather, our observation is simply that the two effects are approximately the same.

SUMMARY AND CONCLUSIONS

We have now looked at some details surrounding the relationship between retention and practice and found that human memory mirrors, with a remarkable degree of fidelity, the structure that exists in the environment. Both display retention and practice functions that are at least approximately power laws. Retention and practice effects are approximately additive. These are not trivial conclusions and other relationships are

---

3. As the Glenberg interaction is basically ordinal, we have chosen to plot Figure 9 in the conceptually simpler need probability rather than the theoretically more correct need odds.

4. Bogartz (1990) focuses on the model in Wickelgren (1972), rather than Wickelgren (1974), where power fits are discussed.
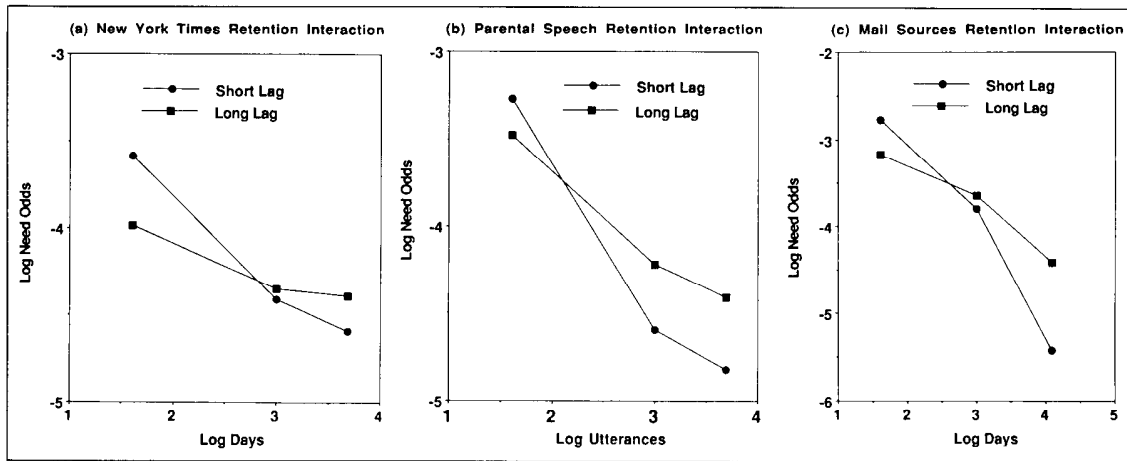
John R. Anderson and Lael J. Schooler

Fig. 9. Retention function for items occurring twice in the previous 100 time units: (a) *New York Times*, (b) child language, and (c) electronic mail messages. Separate functions are plotted for items for whom the two occurrences were less than 10 units apart (short lag) and for whom the two occurrences were at least 10 units apart (long lag).

quite plausible. Evidence for their nontriviality can be seen in the fact that these conclusions have been reached with some reluctance and controversy in psychology—to the extent that we can consider these conclusions established. Finally, there is an interaction between spacing and retention such that retention functions are steeper for more massed practice.

What are we to make of this parallelism between memory and environment? Certainly we can go away with the conclusion that the functioning of memory is remarkably well adapted to the structure of the environment. We also believe that there is a causal link here—that memory has the structure it has because the environment has the structure it has. However, it is possible to hold out for the hypothesis of an accidental correlation between the two.

Formulating the Effects of Practice and Retention

There remains the question of what memory mechanism would actually produce the practice and retention functions we saw. One can aspire to address this question at different levels. One level would be the underlying processes that produce these results. We believe that such an explanation would have to be at the neural level in terms of the physical changes that underlie learning. Short of this, one could aspire to have a mathematical description of how memory would respond to various presentation schedules. There has not been a satisfactory mathematical description to date. However, as a consequence of the analyses we have developed in this paper, we think we are now in possession of such a formulation.
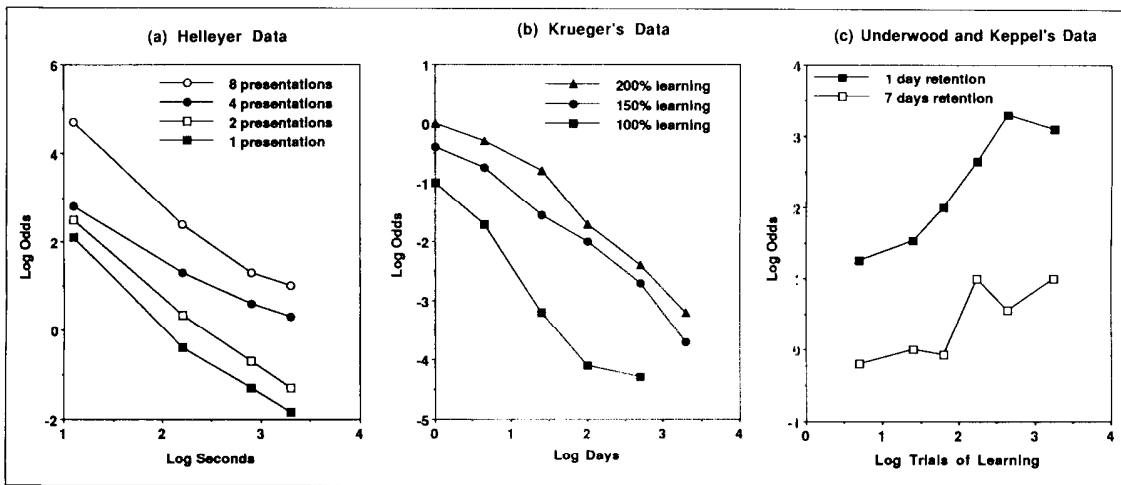

Fig. 10. (a) Forgetting curves at four practice levels from Hellyer (1962); (b) forgetting curves at four practice levels from Krueger (1929); (c) effects of practice at two retention intervals from Underwood and Keppel (1963).

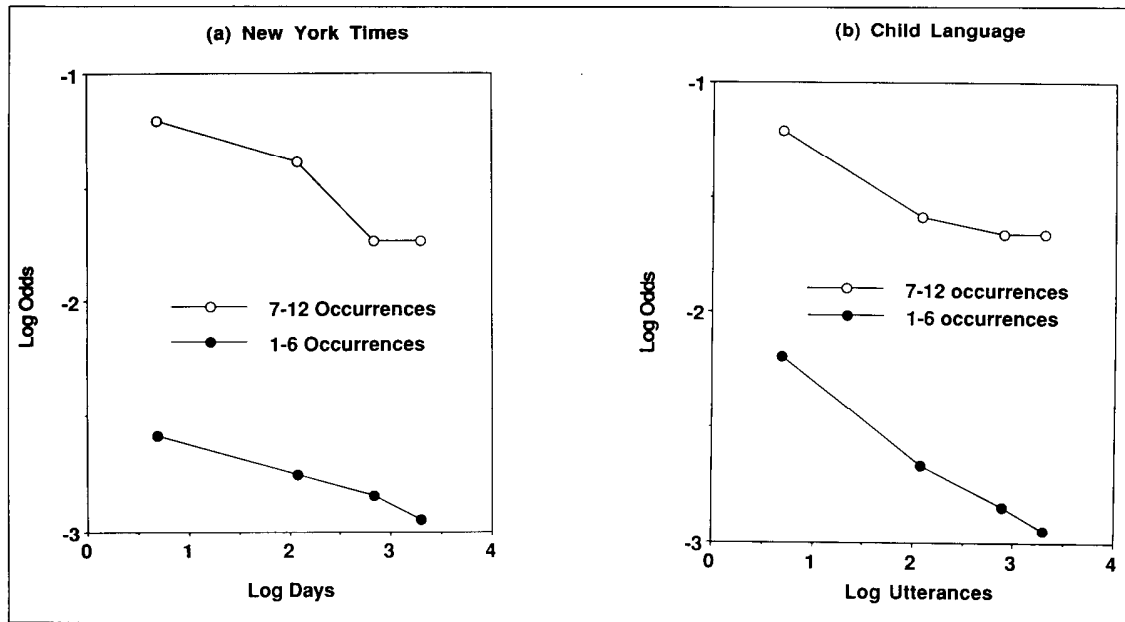Reflections of Environment in Memory

Fig. 11. (a) Retention effects in the *New York Times* for a word with different frequencies of occurrence; (b) retention effects in the child language data base for items with different frequencies of occurrence.

Before providing a mathematical formulation, we would like to state the basic assumptions behind the model:

0. Strength of a trace provides an encoding of its need odds memory performance.

1. The strengths from individual presentations sum to produce a total strength.

2. Strengths of individual presentations decay as a power function of the time.

3. The exponent of the power function for decay of each presentation decreases as a function of time since previous presentation.

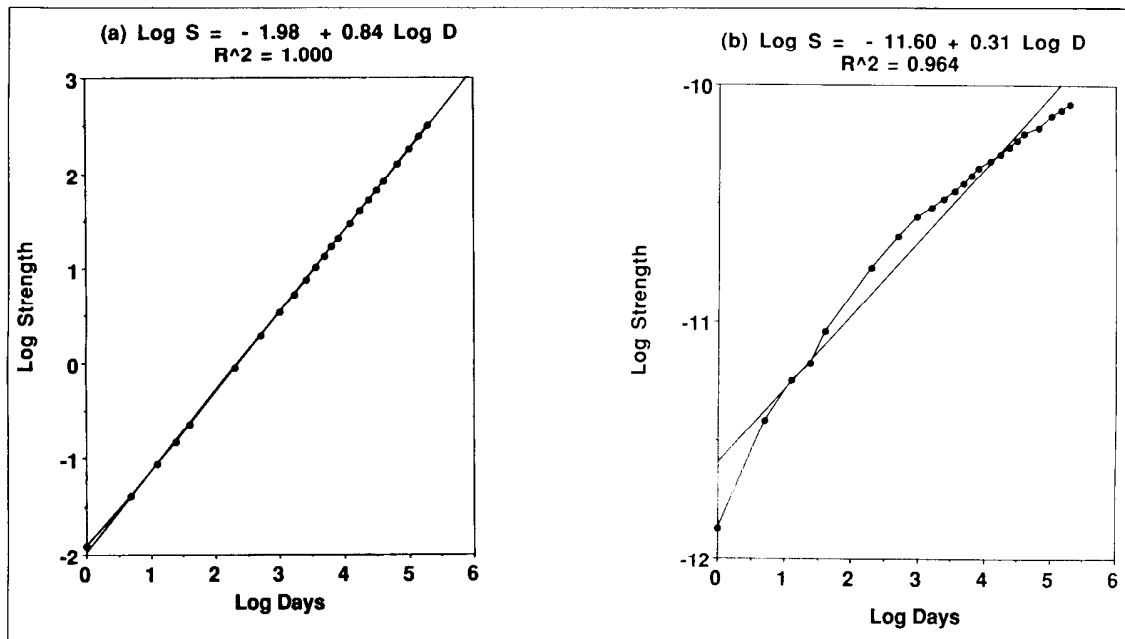We now give an equation to formalize each of the assumptions 1–3.

Fig. 12. Practice functions generated by the mathematical model for (a) $d_1 = .125$ and (b) $d_1 = 1.000$.

John R. Anderson and Lael J. Schooler

Let t_i be the time since the ith presentation of an item and $s(t_i)$ be the strength remaining after this time. Then, corresponding to Assumption 1, we have:

$$S = A \sum_{i=1}^{n} s(t_i) \qquad (9)$$

where S is total strength and A is a scale factor. Corresponding to Assumption 2, we have

$$s(t_i) = t_i^{-d_i} \qquad (10)$$

where d_i is an exponent that can be different for each presentation i. In the case of the first presentation, d_1 is a parameter of the experiment. It may vary with the type of material. Corresponding to Assumption 3, we have for other d_i:

$$d_i = \max[d_1, b(t_i - t_{i-1})^{-d_1}] \qquad (11)$$

that is, d_i is the maximum of the decay rate for the initial presentation d_1, and $b(t_i - t_{i-1})^{-d_1}$. The basic idea is that the decay rate should also decline according to a power function of the time elapsed between the ith and i − 1st presentation, $b(t_i - t_{i-1})^{-d_1}$, but that in no case should it become lower than d_1. Thus, if we wait a short time for a second presentation, the decay rate for the second presentation will be high; whereas if we wait long enough, the decay rate for the second presentation will be no different than for the first. Intuitively, the closer two studies are together, the smaller the contribution of the second makes to the overall strength. While Equation 11 satisfies these constraints, its exact form is a bit arbitrary in that it also has decay rate declining as a power function. There is no evidence one way or the other for this precise an assumption.

We have fit this model to various empirical results. Our goal is to see if we can reproduce the empirical relations we have observed in terms of strength. Since we leave open the mapping of strength onto actual behavioral measures, we can arbitrarily set A = 1 for simplicity. We have also set b = .61, a value that works well for all of our applications, leaving d_1 as the one parameter to be chosen.

The model can obviously produce the phenomena of power-law forgetting, since that is directly built into the retention function. We explored the growth in strength of one practice per day when d_1 = .125 and when d_1 = 1.000. The results are plotted in Figure 12. Both curves approximate a power function quite well, although the approximation is better when d_1 = .125. As can be seen, the exponent of the learning curve decreases with the decay exponent as proposed by Anderson (1982), but it is no longer the case that the two sum to 1.

Next, we investigate whether this model can reproduce the spacing effects. Figure 13 shows the strength calculation for Glenberg's (1976) experiment with d_1 = .125. The correspondence with Figure 4 is compelling. Finally, we attempted a simulation with d_1 = 1.5 of the Hellyer (1962) data in Figure 10a on the additivity of retention and practice effects (Fig. 14). Once again the correspondence between data and simulation is compelling.
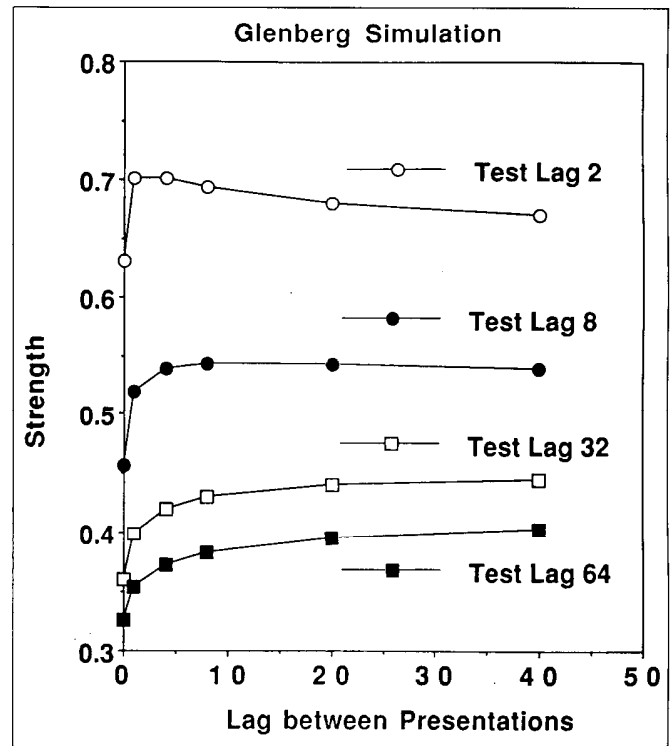

Fig. 13. Simulation of the Glenberg (1976) data (Fig. 4) by the mathematical model.

Although deeper mechanistic explanations would be nice, we think it is an accomplishment to finally have mathematical functions that can capture the effects of practice and delay. We think this has been a direct result of our focus on the structure of the environment. The relationships determining need probability in the environment seem particularly apparent—perhaps because one is not blinded by prior beliefs about mechanistic models.

What Produces the Environmental Structure?

In lieu of a mechanistic explanation, one can ask for an explanation of why the environment displays the relationships it does. Anderson and Milson (1989) can be consulted for the details of an explanation that is an elaborated version of a model proposal by Burrell (1980) to account for library borrowings. It has basically two assumptions. First, it assumes that memories vary in a property called desirability, where a memory's intrinsic desirability determines its rate of use. It turns out that this assumption helps explain frequency and recency effects in that memories that have been used more recently or frequently are more likely under a Bayesian analysis to be highly desirable. Second, the model assumes that memories can rise and fall in this desirability and memories also differ in such volatility. This assumption, again under a Bayesian analysis, helps predict recency and spacing effects. For instance, an item that has had a number of massed presentations a long while ago is identified as probably being a volatile item that had a momentary rise to high desirability and is no longer in use.

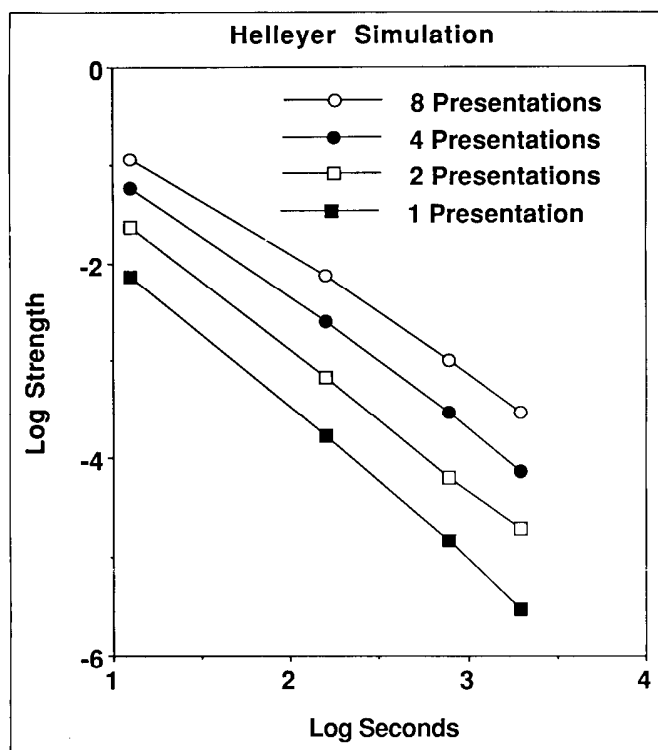## Reflections of Environment in Memory

Fig. 14. Simulation of the Hellyer (1962) data (Fig. 10a) by the mathematical model.

This is not a particularly obscure model of the environmental properties of memories. Nonetheless, it turns out these simple assumptions have led to memory characteristics that have confounded psychologists since Ebbinghaus.

REFERENCES

Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological Review, 89,* 369–406.

Anderson, J.R. (1989). A rational analysis of human memory. In H.L. Roedinger & F.I.M. Craik (Eds.), *Varieties of memory and consciousness* (pp. 195–210). Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J.R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J.R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96(4),* 703–719.

Baddeley, A.D. (1986). *Working memory.* Oxford, UK: Oxford University Press.

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

Bahrick, H.P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113,* 1–24.

Bogartz, R.S. (1990). Evaluating forgetting curves psychologically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 138–148.

Burrell, Q.L. (1980). A simple stochastic model for library loans. *Journal of Documentation, 41,* 100–115.

Ebbinghaus, H. (1964, 1885). *Memory: A contribution to experimental psychology.* Mineola, NY: Dover Publications.

Estes, W.K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377.

Glenberg, A.M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15,* 1–16.

Hellyer, S. (1962). Frequency of stimulus presentation and short-term decrement in recall. *Journal of Experimental Psychology, 64,* 650.

Krueger, W.C.F. (1929). The effects of overlearning on retention. *Journal of Experimental Psychology, 12,* 71–78.

Landauer, T.K. (1975). Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology, 7,* 495–531.

Lewis, C.H. (1978). *Production system models of practice effects.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Loftus, G.R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 397–406.

Logan, G.D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: an update. *Journal of Child Language, 17,* 457–472.

Mazur, J.E., & Hastie, R. (1975). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin, 85,* 1256–1274.

McKay, D.G. (1988). The problem flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review, 89,* 483–506.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Restle, F., & Greeno, J.G. (1970). *Introduction to mathematical psychology.* Reading, MA: Addison-Wesley.

Schooler, L.J., & Anderson, J.R. (unpublished). Environmental demands in memory: Statistical analogs to learning and forgetting curves.

Shepard, R.N. (1990). *Mind sights.* New York: Freeman.

Shrager, J.C., Hogg, T., & Huberman, B.A. (1988). A dynamical theory of the power-law learning in problem-solving. In *Proceedings of the tenth annual conference of the Cognitive Science Society* (pp. 468–474). Hillsdale, NJ: Cognitive Science Society.

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika, 52,* 425–440.

Simon, H.A. (in press). Cognitive architectures and rational analysis: Comment. In K. Van Lehn (Ed.), *Architectures for intelligence.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Simon, H.A., & Ijiri, Y. (1977). *Skew distributions and the sizes of business firms.* New York: Elsevier/North Holland.

Slamecka, N.J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 384–397.

Underwood, B.J., & Keppel, G. (1963). Retention as a function of degree of learning and letter-sequence interference. *Psychological Monographs, 77* (1, Whole No. 567).

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 and Cognition, 2,* 775–780.

Wickelgren, W.A. (1976). Memory storage dynamics. In W.K. Estes (Ed.), *Handbook of learning and cognitive processes* (pp. 321–361). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tuesday, February 11, 2003 (3).max