

THEORETICAL NOTES

On the Form of the Retention Function: Comment on Rubin and Wenzel (1996): A Quantitative Description of Retention

Thomas D. Wickens
University of California, Los Angeles

D. C. Rubin and A. E. Wenzel (1996) fitted many simple functions to a large collection of retention data sets. Their search for the mathematical form of the retention function can be simplified by (a) attending to the failures of simple functions, (b) considering the constraints and process assumptions that any psychological theory must obey, and (c) drawing on results from survival theory. Three sets of psychologically plausible assumptions to interpret the form of a retention function are described. These representations converge on a single functional form, demonstrating the impossibility of determining process purely from empirical fits. A candidate form for an empirical retention function whose parameters separate the various aspects of retention is proposed. These parameters can be used to compare results from different studies.

In their analysis of the form of the retention function, Rubin and Wenzel (1996) search for a mathematical equation to describe the empirical relationship between the time since study and the amount of material retained. (See also related work by Wixted & Ebbesen, 1991, and the earlier work reviewed by Rubin & Wenzel.) They assemble an extensive collection of data sets, each describing an empirical retention function R_i (in itself a substantial contribution). They then fit each set of data with a variety of mathematical functions $r(t)$, their goal being to select the function or set of functions that best characterizes the disparate data sets. They specifically eschew any theoretical basis for their analysis, asserting that a purely empirical approach is more appropriate for the formative stages of psychological theory.

This article points out some limitations to their approach and some ways to simplify their search. I believe that by basing their work entirely on the empirical fit of their functions, and by avoiding what appears to be theorizing, Rubin and Wenzel (1996) have considerably complicated their task. Some minimal consideration of the properties of the forgetting process can both eliminate many candidate functions and suggest others that are more concordant with what must, eventually, be psychological theory. In my discussion, I first look at how the failure of a simple description is often more informative than the fit of a complex description. Next, I turn to the way that knowledge of

the process being described constrains the selection of a function. These considerations lead me to three descriptions of the retention process, each of which implies the same form for $r(t)$, and I briefly describe the interpretation of these functions in terms of the three models. Finally, I combine these observations with some properties of simple stochastic processes, to suggest a candidate for the empirical form of the retention function. Interpretation of the parameters of this model can give psychological insight and certainly brings greater order to the analysis.

The Importance of Failed Models

My first point seems somewhat paradoxical at first. One often obtains more information about a psychological process by looking at how a simple model fails than by finding a complex model that fits. This observation is particularly true when the models are generated without theoretical input.

In their discussion, both Wixted and Ebbesen (1991) and Rubin and Wenzel (1996) emphasize the functions that deviate least from their observed data, and they usually ignore those functions that fail to fit (although Figure 1 of Rubin & Wenzel is a welcome exception).¹ Although some such strategy is neces-

¹ In passing, note that the values of r^2 reported by Rubin and Wenzel (1996) are not Pearson correlations of observed and fitted values, but are the fitted sums of squares relative to sum of squares about the mean, $r^2 = 1 - \frac{\sum [y_i - r(x_i)]^2}{\sum (y_i - \bar{y})^2}$. Correlations are inappropriate here, as they measure the fit of the a linear rescaling of the fitted function, $a + br(t)$, not that of $r(t)$ itself. In effect, using correlation as a measure adds one or two parameters to any function that does not have the form $a + bf(t)$ for some $f(t)$. The difference between their r^2 and a correlation can be substantial—for example, the exponential function $e^{-\alpha t}$ in Figure 1 actually explains less variability than the mean alone, yet has a squared correlation of .85, which measures the fit of the function $a + be^{-\alpha t}$.

This work was supported in part by a University of California, Los Angeles (UCLA) Faculty Research Grant. I thank Eric W. Holman and the members of the Cogfog group for useful discussion.

Correspondence concerning this article should be addressed to Thomas D. Wickens, Department of Psychology, UCLA, Box 951563, Los Angeles, California 90095-1563. Electronic mail may be sent to twickens@psych.ucla.edu.

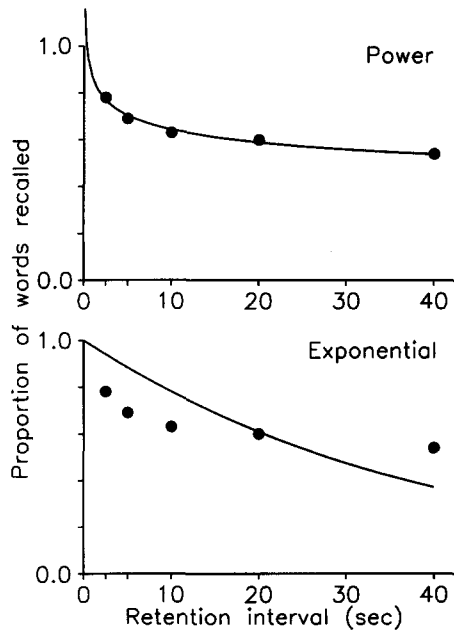


Figure 1. A power function $r(t) = 0.86t^{-0.13}$ and an exponential function $r(t) = e^{-0.025t}$ fitted to the data from the 5-s group in Wixted and Ebbesen's (1991) Experiment 1.

sary when one searches among such a large set of candidates, as do Rubin and Wenzel, both empirical and theoretical insight are gained by looking at how the data deviate from a conceptually simple representation. The analysis of Wixted and Ebbesen provides a good example. They report two retention experiments and note (in their Table 2) that a power function fits better than do five other functions. They show that an exponential function is unsatisfactory, but they do not pursue the point. Because they emphasize obtaining a good fit, their illustration (their Figure 1) uses a power function. However, in many respects, a plot of the exponential function is more informative than that of the power function.

The exponential function,

$$r(t) = be^{-ct}, \quad (1)$$

has a long history and many important and well-known properties. It is characterized by two parameters, an initial level b and a rate of drop c . Among its many interpretations, it describes the size of a population of independent entities, each of which has a constant probability of decaying (for example, see Wickens, 1982, Section 10.1 or practically any book on stochastic processes or queuing theory). When used as a retention function, it suggests that each member of a set of items becomes unavailable with a constant likelihood in each short interval. This simple model, known as a *Poisson death process*, provides a baseline against which to compare an observed function.

Specifically consider the data from the 5-s group in Wixted and Ebbesen's (1991) Experiment 1. Figure 1 shows these data fitted by a power function and by an exponential function with $b = 1$ and c free.² The upper panel tells us only that the power function fits well. The lower panel shows both that the one-

parameter exponential function fails and that it fails because the exponential function overestimates points at short retention times and underestimates them at long retention times. The observed retention function falls more rapidly than the exponential early in its course and more slowly than the exponential later on. This observation immediately suggests three descriptions of the forgetting process, each based on the way that a collection of distinct items become inaccessible:

1. *Heterogeneity.* The observed retention function arises from a mixture of easy and hard items. The loss of items is described by a collection of independent Poisson death processes with a distribution of rate parameters. Retention drops quickly at first as the fragile processes perish, then more slowly when only the more robust processes remain.³
2. *Consolidation.* An item gains strength throughout the time that it remains available, perhaps through some form of covert rehearsal or through integration into larger cognitive structures. Although the individual processes are homogeneous and independent, the rate parameter c increases with t . Retention drops quickly at first when all items are weak, then more slowly as items gain strength.
3. *Competition.* The unforgotten items compete, perhaps for shared resources, such as time for rehearsal or connection to some shared list marker or node. Although the individual processes are homogeneous, their rate parameter c depends on the proportion of the other items that remain available. Retention drops quickly at first, when many items complete, then more slowly when there is little competition.

These representations are not novel; I drew them from standard techniques in the theory of birth and death processes. I have two reasons for raising them. First, they show how the failure of the exponential model suggests psychological process. Second, they imply specific forms for the retention function, which I will discuss below.

Constraints and Boundary Conditions

Now consider the collection of functions that might be fitted to a set of data. Among the many possibilities, some are consistent with the phenomenon being studied, whereas others, possibly outside the range of the data, imply behavior that is impossible.⁴ I feel considerably stronger about the importance of these conditions than do Rubin and Wenzel (1996) (see their remarks on page 749 and their discussion of particular functions on

² I took my data from Wixted and Ebbesen's (1991) figure, so my parameter values differ slightly from theirs. Estimating b improves the fit and makes the discrepancy less obvious but does not change the conclusions.

³ This explanation is considered in detail with respect to these data by Anderson and Tweney (1997), Wixted and Ebbesen (1997).

⁴ I am reminded of being puzzled in my youth on seeing a graph of some of my father's galvanic skin resistance conditioning data that had been fitted with a fourth-degree polynomial by a zealous student of mathematics (Wickens, Gehman, & Sullivan, 1959). Although this function passed closely through the observed data points, it fluctuated wildly out of the plausible range of the dependent variable between points and diverged rapidly outside the range of the independent variable. Similarly, a fourth-degree polynomial perfectly fits the data in Figure 1, but is completely useless.

pages 745–749). I believe that it makes little sense to fit a function that makes incompatible predictions. At best, this function approximates a more accurate representation. With the wealth of simple mathematical functions that are consistent with retention data, it is foolish to expend much effort fitting inconsistent functions. The constraints are particularly important if one wishes to think of the retention functions as extensible to a wider range of retention times (Rubin & Wenzel suggest that a 90-to-1 range is appropriate). There is no reason to consider functions that would be rejected, were such data available.

An analysis of the forgetting process suggests three considerations, two boundary conditions and one restriction on form, that apply to the retention function:

1. *The initial level.* Each measure of retention has a value or range of values that is logically consistent with the state of the memory at time $t = 0$. For example, the proportion of correct retrievals cannot be outside the unit interval; the logit of the probability can be any real number including $+\infty$, and a d' statistic should be large, but not infinite. This constraint eliminates many functions. For example, one cannot with any consistency fit proportion data by a function that goes to infinity as t goes to zero. Thus, the simple power function fitted in the top panel of Figure 1 is unsatisfactory (notice that it pokes above one).
2. *The final asymptote.* As with the initial level, each measure has a minimum value to which it falls when all mnemonic information is extinguished. For correct-response proportions this value is either zero or some guessing probability such as the reciprocal of the number of response alternatives; for d' it is zero; and for the probability of a correct response in a forced-choice recognition task it is one half. Functions that drop below this level should not be fitted. A nonzero asymptote is acceptable if some form of permanent storage is postulated.
3. *Monotonicity.* In the designs that characteristically are used to determine forgetting functions, the measure of memory declines with the delay between study and test.⁵ Thus, any satisfactory function should decrease monotonically with t . More formally the first derivative $r'(t)$ of $r(t)$ should never be positive when $t \geq 0$. This constraint eliminates most functions based on polynomials. For example, Rubin and Wenzel (1996) use this criterion to eliminate the function $r(t) = (b - mv\sqrt{t})^2$ (it also violates my asymptote condition).

These criteria restrict both the possible functional forms and the values that the parameters can take. For example, the exponential function $r(t) = a + be^{-ct}$ only satisfies the three criteria when $0 \leq a \leq 1$, $0 \leq b \leq 1 - a$, and $c > 0$. If psychological insight is to be obtained from the form of the retention function, then functions that are inconsistent with these criteria should not be considered, regardless of how well they match particular sets of data.

As these constraints show, the set of plausible functions depends in part on how retention is measured. For example, a function that is satisfactory as a representation of retention measured by d' must be able to start above unity, a property that violates the boundary conditions for retention measured by retrieval probability. Which measure was used must be considered as part of the fitting procedure.

How $r(t)$ is measured has another implication for the fitting of functional forms. What constitutes a satisfactory form for one measure is not the same as that for another. Some measures, such as probabilities, odds ratios, and logits, are nonlinear math-

ematical functions of each other. These are easily converted, one to another, and a particular functional form for one measure implies a different form for the other. For example, if the probability of a correct response can be fitted by an exponential function, then the odds can be fitted by the function $r(t) = (b' - 1)^{-1}$, but not by an exponential. A translation between measures that are not connected by a mathematical relationship is more difficult. Conversion between retention measured by probabilities, d' , and the latency of retrieval cannot be done atheoretically. Some form of model of the retention process is necessary. Whatever the assumptions of this model, the relationship between measures will not be linear, so that again different functional forms would be expected to fit. Because a function form for one measure implies a different form for other measures, it is unreasonable to expect that data measured in disparate ways will be fitted by a single best function.

Forgetting Rates and Hazard Functions

Although a study of the relationships among the retention measures is necessary for a full analysis of the retention function, the issue is tangential to my arguments here. To show how an analytic treatment aids the evaluation process, I will use the probability of a correct recall. Working with this measure gives access to the well-developed, but still relatively atheoretical, techniques from survival analysis.

As it refers to data, the retention function $r(t)$ describes the relationship between the delay at which memory is probed and the proportion of material that is retained. However, the equation for $r(t)$ is not the easiest form in which to relate the characteristics of the function to the underlying forgetting events. A retention measure more closely related to the process of forgetting is the *hazard function*, which is the rate at which items are being forgotten relative to how much is still remembered. The rate of retention drop is measured by the derivative $r'(t)$. With retention functions, it is somewhat more natural to speak of the forgetting rate $-r'(t)$ instead of the derivative $r'(t)$. Where $-r'(t)$ is large, much material is being forgotten; and where it is small, little is being lost. However, in a 100-item list, a loss of 3 items has different implications early in the retention period, when 80 of those items can be remembered, than it does late in the retention period when only 6 items are available. To put the derivative on a relative basis, divide it by the opportunities to forget, that is, by the value of $r(t)$, to give the hazard function⁶:

$$h(t) = \frac{-r'(t)}{r(t)}. \quad (2)$$

The hazard function $h(t)$ expresses the instantaneous rate of forgetting of an item that has been retained to time t . It can rise or fall with t ; the only constraint is that it cannot be negative.

⁵ Imposing this restriction does not deny the possibility that in some designs the amount of material recovered may reliably increase between certain time points. It does indicate that such hypermnesic increases are interesting because they are exceptions to the typical pattern.

⁶ When referring to a distribution of lifetimes with density function $f(t)$, the hazard function is $h(t) = f(t)/[1 - F(t)]$, a form found in many statistical references. As Equation 2 shows, it is also a logarithmic derivative: $h(t) = -d \log r(t)/dt$.

The hazard functions of several candidates for the retention function are given in Table 1. Three of these functions are plotted in Figure 2, for parameter values consistent with Wixted and Ebbesen's (1991) Experiment 1.

The simplest among the hazard functions is that of the exponential function with zero asymptote. It is a constant, independent both of the initial level of retention and the value of t . It is the benchmark to which the other hazard functions are compared. The hazard function for every other distribution changes with t and thus is above that of an exponential function with equivalent rate in some places and below it in others. The observation that Wixted and Ebbesen's (1991) Experiment 1 data in Figure 1 fall more rapidly than an exponential at first and more slowly later implies that items are at greater risk early in the retention process and that the hazard function falls.

As an analysis tool, the hazard function is more suited to conceptual analysis than to empirical testing. I use it in the next section to generate several retention functions, but I have not plotted any data in Figure 2. Although the use of empirical estimates of the hazard function to select among the candidate functions or explanations is attractive, adequate precision is hard to obtain, particularly from studies not designed for the purpose. The hazard function $h(T)$ at time T is estimated by the ratio of an estimate $\hat{r}'(T)$ of the derivative $r'(T)$ to an estimate $\hat{r}(T)$ of $r(T)$. Denote the observed retention at time t by R_t , and suppose that retention has also been measured at times ΔT before and after T . The hazard function is estimated by the ratio of the slope between the two outer points to the retention at the midpoint:

$$\hat{h}(T) = \frac{-\hat{r}'(T)}{\hat{r}(T)} = \frac{R_{T-\Delta T} - R_{T+\Delta T}}{2\Delta TR_T}$$

However, a researcher using this formula is encumbered by the difficulty of estimating the derivative: when ΔT is too large, $r'(t)$ changes between the three points and the estimate is inaccurate; when ΔT is too small, the difference $R_{T-\Delta T} - R_{T+\Delta T}$ is dominated by sampling error. The problem is exacerbated in the right tail of the function, where $r(t)$ is small and the uncertainty in R_T is a large proportion of its value.⁷ The only really stable

Table 1
Slope and Hazard Functions of Some Common Retention Functions

Function	$r(t)$	$-r'(t)$	$h(t)$ or $h(r)$
Linear	$b - ct$	c	$\frac{c}{b - ct} = \frac{c}{r}$
Exponential	be^{-ct}	bce^{-ct}	c
Hyperbolic	$\frac{b}{1 + ct}$	$\frac{bc}{(1 + ct)^2}$	$\frac{c}{1 + ct} = \frac{c}{b} r$
Power	bt^{-d}	$bd t^{-(d+1)}$	$\frac{d}{t} = \frac{d}{b^{1/d}} r^{1/d}$
Pareto-II	$\frac{b}{(1 + ct)^d}$	$\frac{bcd}{(1 + ct)^{d+1}}$	$\frac{cd}{1 + ct} c^{d/\sqrt{r/b}}$
Weibull	$b \exp[-(ct)^d]$	$bcd \exp[-(ct)^d](ct)^{d-1}$	$cd(ct)^{d-1}$

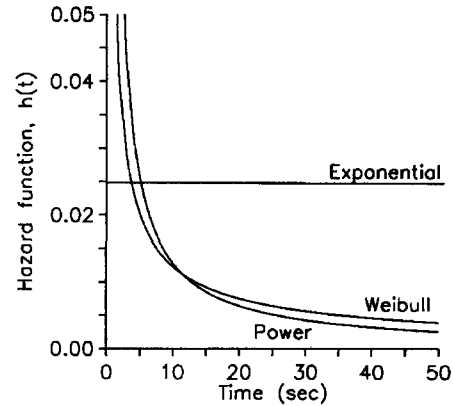


Figure 2. Hazard functions for functions fitted to data from the 5-s group in Wixted and Ebbesen's (1991) Experiment 1.

way to estimate the hazard, particularly near the extreme of the function, is to smooth the retention data by fitting a function to it, then use the hazard function of the fitted function as an estimate of the actual function. In simple form, this strategy is equivalent to fitting the retention function.

Three Nonexponential Models

Earlier I proposed three interpretations of the failure of an exponential function to fit retention data. I now develop corresponding forms for $r(t)$. The analysis illustrates three points. First, a simple theoretical analysis can lead to a functional form that is easily overlooked, even in a compendium as large as that of Rubin and Wenzel (1996). Second, different theoretical explanations can imply identical functional forms, limiting the utility of a purely empirical fit to distinguish among theoretical explanations. Finally, each interpretation yields its own information about the memory system, which can be expressed as an appropriate descriptive plot. Each gives its own characterization of how the retention function differs from a simple exponential.

I will illustrate these analyses with the two sets of data used as examples by Rubin and Wenzel (1996) (Figure 3, taken from their Figure 1). They obtained one set by averaging the results from Peterson and Peterson (1959) and Murdock (1961) for short-term retention of letter or word trigrams. An exponential function fits these data fairly well. They obtained the other set by averaging long-term memory for campus locations in six studies reported by Bahrack (1983). For these data, an exponential function fails badly.

First consider the model of heterogeneous items. A venerable approach to variability extends the single-item model by treating the recorded data as a mixture of items (see Wickens, 1982, Section 3.5). Each item has an exponential retention function $r_c(t)$ with $b = 1$. The retention rate c is a random variable, with a distribution having density $f(c)$. This distribution could arise in many ways, such as by pooling over subjects with disparate mnemonic abilities, by pooling over items with disparate mem-

⁷ Similar difficulties in determining the hazard function for response times are discussed by Luce (1986, Section 4.1.1).

orability, by pooling across a set of distinct internal representations of the item, or (most likely) by a combination of these possibilities. The composite retention function is the average (over c) of its individual components:

$$r(t) = E_c[r_c(t)] = \int_0^\infty r_c(t)f(c)dc. \quad (3)$$

This mixture of exponentials has itself exponential form only when the distribution of c is degenerate, with all mass at one value.⁸

This description displaces the problem of finding the form of the retention function to that of finding the distribution of forgetting rates c . Certain distributions give $r(t)$ convenient forms. One natural choice is the gamma distribution (e.g., Wickens, 1982, Section A.4). This distribution depends on two parameters, α and β , and has the density function

$$f(c) = \frac{\alpha^\beta}{\Gamma(\beta)} c^{\beta-1} e^{-\alpha c}. \quad (4)$$

The parameter β determines the shape of this distribution, which is exponential when $\beta = 1$ and becomes more symmetrical as β increases. The parameter α scales the distribution such that its mean is β/α . Substituting the gamma density into Equation 3 and evaluating the integral gives the retention function

$$r(t) = (1 + t/\alpha)^{-\beta}.$$

For simplicity and consistency with later equations, reparameterize this result by putting $c = 1/\alpha$ and $d = \beta$ to give the form

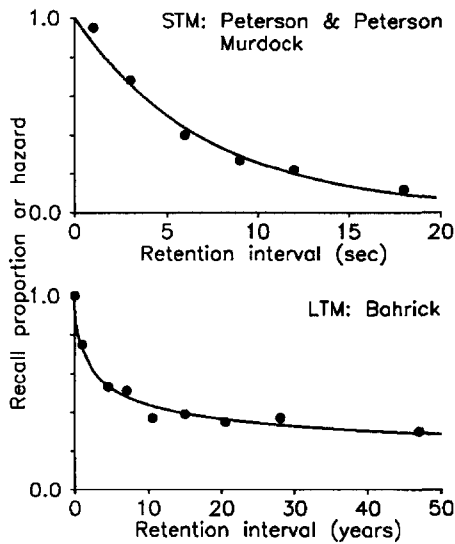


Figure 3. Retention data for short-term memory (STM) and long-term memory (LTM) from Rubin and Wenzel (1996), with fitted Pareto-II distributions and corresponding hazard functions. Short-term memory data were obtained by averaging results from Peterson and Peterson (1959) and Murdock (1961). Long-term data were obtained by averaging results for six studies reported by Bahrick (1983).

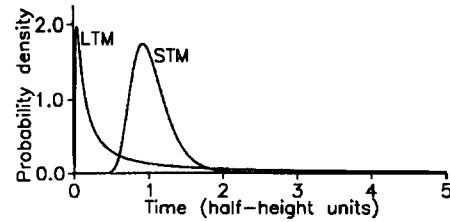


Figure 4. Distribution of half-lives for heterogeneous exponential processes that produce the Pareto-II functions in Figure 3. The abscissa is scaled by half-height times for $r(t)$. STM = short-term memory; LTM = long-term memory.

$$r(t) = (1 + ct)^{-d}. \quad (5)$$

The distribution of retention times implied by Equation 5 is a Pareto distribution of the second kind (Johnson, Kotz, & Balakrishnan, 1994, Chapter 20). Its derivation as a gamma mixture of exponentials dates back at least to 1952 (Maguire, Pearson, & Wynn, 1952). It was not in the set examined by Rubin and Wenzel (1996), showing the difficulties of selecting the most informative functions on purely empirical grounds. However, except for the unit offset added to ct , it is similar to the power function $r(t) = bt^{-d}$, which they examined and found satisfactory. One would expect the Pareto-II function to fit about as well.

The Pareto-II function fits the two sets of observations in Figure 3. The parameter estimates are $\hat{c} = 0.007948$ and $\hat{d} = 17.51$ for the short-term function, explaining 98% of the variability, and $\hat{c} = 2.046$ and $\hat{d} = 0.2698$, explaining 96%, for the long-term function. These retention functions are plotted in Figure 3. When expressed as parameters of a gamma mixing distribution, these estimates are $\hat{\alpha} = 1/\hat{c} = 126.6$ and $\hat{\beta} = \hat{d} = 17.51$ for the short-term function and $\hat{\alpha} = 0.488$ and $\hat{\beta} = 0.270$ for the long-term function.

A plot of the item distributions shows how the two sets of data differ under the mixture model. Obviously, the most salient aspect of their difference is their time scale. To allow them to be plotted on the same display, I have expressed them in terms of the time taken to fall to half their height, that is, so that $r(1) = 1/2$. Thus, unity corresponds to 5.08 s for the short-term function and 5.89 years for the long-term function. Figure 4 plots the distribution of items under this rescaling, expressed as half-lives of the exponential processes.⁹ The difference between the two data sets is striking. Almost all the short-term items have half-lives between $1/2$ and 2 times the process's half-height time, while the modal long-term item has a half-life that is but

⁸ To see that this is the only possibility, note that with exponential $r_c(t)$, the right-hand side of Equation 3 is the moment-generating function $M_c(-t)$ of the distribution of c . If $r(t)$ is exponential, then the left-hand side is $r(t) = e^{-\beta t}$, and the equation becomes $M_c(t) = e^{\beta t}$. This is the moment-generating function of a degenerate distribution with all mass at β .

⁹ Standardization to half-height times amounts to setting $c = 2^{1/d} - 1$. Figure 4 plots the distribution of $T = 1/(kc)$, where $k = \log 2$, which has density $g(t) = k(kt)^{-2}f_c[1/kt]$.

a tiny fraction of the half-height time. These quickly decaying items are balanced by a very long right tail of the distribution, an effect that would be even more striking if the abscissa were extended to the full duration of the data at a bit over eight half-height times.

Both the consolidation and competition interpretation of non-exponential retention are best approached through the hazard function. Somewhat surprisingly, these models can also give Equation 5. For the consolidation model of acquired resistance to loss, assume that the items are independent and homogeneous and that their hazard function starts at α when $t = 0$ and declines with t as surviving items accumulate strength. Their resistance to loss grows without bound, so $h(t) \rightarrow 0$ as $t \rightarrow \infty$. One way to write such a function is as an offset reciprocal:

$$h(t) = \alpha / (1 + \beta t).$$

This function is identifiable from Table 1 as the hazard function of the Pareto-II distribution.¹⁰ The two data sets under this interpretation are represented by plotting their hazard functions with the abscissa scaled by the half-height time (Figure 5). The different character of the two retention functions is obvious. The short-term hazard function is nearly flat, indicating that no consolidation takes place. The long-term function falls sharply and shows considerable consolidation.

To obtain the retention function for the model of competition among items, again assume that the processes are homogeneous with a hazard function that starts at $h(0) = \alpha$. Competition acts on the individual items, so is described by its effects on the hazard function. From its initial value of α , when all items are available ($t = 0$ and $r = 1$), it declines as items are lost and competition decreases. A simple choice for this for this relationship is a power function with exponent $\beta > 0$:

$$h(r) = \alpha r^\beta.$$

This function again is recognizable from Table 1 as that of a Pareto-II distribution.¹¹ This model is characterized graphically in Figure 6 by plotting the hazard as a function of the proportion of surviving items. Note that the early states of the process are on the right of this plot and the late states are on the left. Once again, the difference between the functions is clear. The short-term hazard is almost independent of r , while the long-term hazard is strongly related to r .

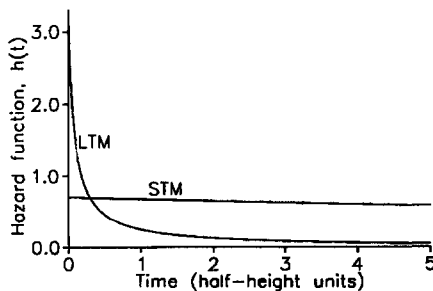


Figure 5. Pareto-II hazard functions for the data in Figure 3. The abscissa is scaled by half-height times for $r(t)$. STM = short-term memory; LTM = long-term memory.

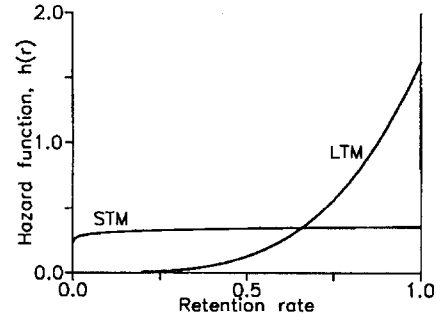


Figure 6. Pareto-II hazard functions for the data in Figure 3 plotted as functions of the recall rate. STM = short-term memory; LTM = long-term memory.

I do not wish to suggest that the convergence of these three models to the Pareto-II form gives that function some unique or particularly desirable status. Other choices for the mixing distribution or the hazard function, consistent with the same basic assumptions, give different final forms. What is noteworthy is that quite different sets of process assumptions give the same functional form. This fact illustrates the difficulties involved in trying to infer process from the empirical form of the retention function. Knowing that the Pareto-II retention function fits a set of data well does not help one to decide among the psychological interpretations. This ambiguity is common: Simple function forms observed in data rarely, if ever, constrain the range of plausible psychological processes sufficiently definitively to be useful by themselves. I do not view these limits pessimistically, however. The use of the Pareto-II distribution is far less important than the way that it allows the differences between the short-term and long-term data to be characterized in Figures 4–6. Each interpretation suggest a way to plot the results, and each suggest particular experimental manipulations that should, if the causes are as hypothesized, affect retention in characteristic ways.

Characteristics of Forgetting Functions

A valuable contribution of the Pareto-II representation is the way that it allows the differences between the short-term and long-term data to be characterized by parameter values. Parameter c expressed the time scale of the process and d the deviation from exponential form. In this descriptive spirit, we can try to develop a prototype for the retention function that has the flexibility to fit a variety of data sets and that isolates the important characteristics of the forgetting process in distinct parameters. Four properties are reasonable to include in such a function:

1. *The initial level.* Each function starts with an initial level of learning. As discussed above, this level is constrained by the

¹⁰ It can also be found by integrating the differential equation in Footnote 6: $\log r(t) = \int h(t)dt = -(\alpha/\beta) \log(1 + \beta t) + C$, which, after exponentiation and reparameterization, gives Equation 5.

¹¹ To obtain this result analytically, rearrange Equation 2, then integrate: $t = \int dt = -\int dr/[rh(r)] = 1/(\alpha\beta r^\beta) + C$, which, after solving for $r(t)$ and reparameterizing, gives Equation 5.

measure used, but it can also be a quantity of theoretical interest, particularly when using d' statistics or with weakly learned material. The initial level is readily accommodated by a multiplicative parameter:

$$r(t) = bf(t),$$

where $f(0) = 1$, and the function $f(t)$ remains otherwise to be determined. For the probability of retention, $0 < b \leq 1$.

2. *The final asymptote.* Each retention function is also characterized by its limit as $t \rightarrow \infty$. The value of $r(t)$ at this limit is a critical component of a theory of forgetting, for it distinguishes between complete lability of the trace and some form of permanent store. The final asymptote is represented by an additive parameter, so that

$$r(t) = a + bf(t),$$

with $f(t) \rightarrow 1$ as $t \rightarrow \infty$. Together the parameters a and b determine the left and right ends of the retention function, and $f(t)$ is a cumulative distribution function. For the probability of retention, $0 \leq a < 1$ and $0 < b \leq 1 - a$.

3. *The rate of loss.* Each retention function is characterized by something like the average rate at which material is lost. The rate of loss can be accommodated by rescaling the time dimension. A simple multiplicative rescaling gives

$$r(t) = a + bf(ct).$$

Large values of c imply that the material is quickly forgotten. Every function in Table 1 has this form, although c is not identifiable for the power function.

4. *Deviations from constant hazard.* When compared to the constant-rate exponential retention function, most simple hazard functions are either high early in the time course of retention and low in the later parts, or low at first and high later on. There is no single natural representation of this effect, and the most useful parameterization undoubtedly awaits a theoretical analysis. One possibility is the parameter d of the Pareto-II function used above. Another approach is to adopt a Weibull (or "exponential-power") function, which has been widely used for this purpose (Johnson et al., 1994, Chapter 21, give 28 pages of references to this distribution and its applications; Rubin & Wenzel, 1996, give further citations). Incorporating this function into the retention model gives

$$r(t) = a + b \exp[-(ct)^d]. \quad (6)$$

Processes with $d = 1$ are exponential, those with $d > 1$ have an increasing hazard function, and those with $d < 1$ have a decreasing hazard function.¹²

The four parameters of Equation 6 are too many to expect to extract from a retention function based on only a few points. As Rubin and Wenzel (1996) point out, the typical study does not report data from enough time points to discriminate among the more complex functional forms. In practice, researchers measuring the course of forgetting have emphasized one or two of the parameters and ignored or assumed values for the others. Appeal to the boundary conditions mentioned above helps. The value of a is difficult to determine experimentally and usually risky to extrapolate to from a fitted function, so complete forget-

ting at long times is often assumed by setting $a = 0$. It is also often convenient to assume that retention at $t = 0$, if measurable, would be perfect, at least with well-learned material. Together these assumptions reduce Equation 6 to a two-parameter Weibull function that is a good starting point for the simple probability of retention:

$$r(t) = \exp[-(ct)^d]. \quad (7)$$

This function fits the Figure 1 data well, with $\hat{c} = 0.00526$, $\hat{d} = 0.289$, and the ability to explain 95% of the variability. Rubin and Wenzel (1996) do not fit either Equation 6 or Equation 7, but they do approximately fit a three-parameter version of Equation 6 with $a = 0$ by stepping d in half units from -2 to 3 in their selection of functions and estimating b and c . When $d = 1/2$, this function was one of their most satisfactory alternatives.

Many characterizations of the forgetting process have emphasized only the rate of loss of the learned material, embodied in the parameter c . This attention is natural and proper, as the overall rate of forgetting, or some surrogate of it, is usually the most obvious characteristic of the forgetting function. It is the parameter that is most immediately affected by most experimental manipulations. However, as a characterization of the forgetting process itself, the Weibull exponent is the most interesting parameter. More than the other parameters, it separates different classes of forgetting process, distinct from such things as the amount of original learning, the interference from other sources, and the rate of forgetting. It captures the difference between the two functions in Figure 3 clearly. For the short-term data, for which the exponential function fit fairly well, $\hat{d} = 1.05 \approx 1$. For the long-term data, for which the exponential fails, $\hat{d} = 0.34 < 1$. As the discussion of the three Pareto-II representations shows, it is impossible to tell what aspect of the two studies produced this difference without further results. It will be interesting to compare these estimates with those obtained from other short-term and long-term retention studies.

This analysis suggests how to follow up Rubin and Wenzel's (1996) treatment of their wide collection of data. First fit a model such as Equation 7 to the individual data sets, after expressing them as some standard measure (for example, converting odds ratios, logits, and d' values to correct-response probabilities). Take note of serious failures to fit, although without concern about small differences in goodness of fit. For the data successfully fitted, estimate the parameters and examine how these values relate to the characteristics of the individual experiments. Note how procedural differences are reflected in

¹² Perhaps the most intriguing aspect of the Weibull function from a modeling perspective is its relationship to the limiting form of an extreme-value distribution. For a substantial class of parent distributions, the standardized distribution of the minimum of N observations tends to Weibull form as N increases, rather in the way that the standardized sum of a collection of N observations converges to a normal distribution as N increases (Johnson, Kotz, & Balakrishnan, 1995, Chapter 21; Stewart & Ord, 1994, Sections 14.13-17). Several of Rubin and Wenzel's (1996) interpretations derive from this fact. The standardization is necessary here: The distribution is not simply an asymptotic maximum (e.g., see Logan, 1995, and Colonius, 1995).

the parameters. Attention to the parameter values is a necessary part of the procedure. Models such as Equations 7 and 6 are very flexible, and data with quite different characteristics can satisfy the same functional form. For example, data with a rising hazard function ($d > 1$ in Equation 7) are very different from those with a falling hazard function ($d < 1$), although both may be fitted by the same Weibull form.

The use of a single functional form is critical to this approach, which emphasizes fitting the data sets to an interpretable function over finding a "best" function. Any substantial failures to fit should be examined in the context of the originally fitted function, not by searching for new forms. An approach like that I used with the exponential model in Figure 1 is useful here. Look at how the data deviate from the fitted function to see why the failure occurs. Problems associated with the initial or final states of the process can be rectified by adopting a three-parameter version of Equation 6. These extensions will probably accommodate most difficulties. In principle, non-Weibull characteristics of the hazard function, such as nonmonotonicities, could also be at fault, although I suspect that most data sets contain too few time points to resolve these effects.

In a sense, this strategy reverses the one proposed by Rubin and Wenzel (1996). Instead of looking at a wide class of functions, without attention to the parameters of these functions, apply a flexible, but consistent, representation to every data set and see how they differ quantitatively. If a purely empirical approach to the retention function is to be fruitful, surely this is the next step.

References

- Anderson, R. C., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory and Cognition*, 25, 724–730.
- Bahrack, H. P. (1983). The cognitive map of a city: Fifty years of learning and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 17, pp. 125–163). New York: Academic Press.
- Colonus, H. (1995). The instance theory of automaticity: Why the Weibull? *Psychological Review*, 102, 744–750.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions* (Vol. 1, 2nd ed.). New York: Wiley.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2, 2nd ed.). New York: Wiley.
- Logan, G. D. (1995). The Weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review*, 102, 751–756.
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental operations*. New York: Oxford University Press.
- Maguire, B. A., Pearson, E. S., & Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika*, 39, 168–180.
- Murdock, B. B., Jr. (1961). The retention of individual items. *Journal of Experimental Psychology*, 62, 618–625.
- Peterson, L. R., & Peterson, M. J. (1959). Short term retention of individual verbal items. *Journal of Experimental Psychology*, 58, 193–198.
- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103, 734–760.
- Stewart, A., & Ord, J. K. (1994). *Kendall's advanced theory of statistics: Distribution theory* (Vol. 1, 6th ed.). London: Arnold.
- Wickens, D. D., Gehman, R. S., & Sullivan, S. N. (1959). The effect of differential onset time on the conditioned response strength to elements of a stimulus complex. *Journal of Experimental Psychology*, 58, 85–93.
- Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco, CA: Freeman.
- Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2, 409–415.
- Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory and Cognition*, 25, 731–739.

Received August 19, 1996

Revision received June 4, 1997

Accepted June 9, 1997 ■