Journal of Experimental Psychology:
Learning, Memory, and Cognition
1999, Vol. 25, No. 5, 1161-1176

0278-7393/99/$3.00

The Precise Time Course of Retention

David C. Rubin and Sean Hinton
Duke University

Amy Wenzel
University of Iowa

Fits of retention data were examined from 5 conditions: 3 types of cued recall, an old–new recognition task, and a remember–know recognition task. In each condition, 100 participants had either 18 recall or 27 recognition trials at each of 10 delays between 0 and 99 intervening items, providing the first data obtained in experimental psychology that were precise enough to distinguish clearly among simple functions. None of the 105 2-parameter functions tested produced adequate fits to the data. The function $y = a_1e^{-t/1.15} + a_2e^{-t/T_2} + a_3$ fit each of the 5 retention conditions. The T_2 parameter in this equation equaled 28 for the 3 recall conditions and the remember–know recognition condition and 13 for the old–new recognition condition. Individuals' recall data fit the same function with parameters varying with gender and scholastic aptitude scores. Reaction times support the claim that the $a_1e^{-t/1.15}$ term describes working memory, and the remaining 2 terms describe long-term memory.

The goal of mathematically describing retention is as old as the experimental study of memory (Ebbinghaus, 1885/1964), yet no data exist that are precise enough to allow discrimination among the different mathematical functions commonly proposed. Rubin and Wenzel (1996) reviewed the substantial literature on existing retention functions. They found over 200 data sets in the literature and fit them all to 105 two-parameter functions. The data sets included the best available: All had 5 or more retention intervals and were smooth enough to correlate with at least one function .9 or greater. The following could be fit to four functions: recall, recognition, and sensorimotor tasks in people with retention intervals ranging from seconds to decades; delayed matching to sample in birds, rodents, and primates; and all other procedures and species, except autobiographical memory tasks. These functions were the logarithmic, $y = b - m \cdot ln(t)$ (favored by Woodworth, 1938, and other early researchers); the power, $y = b \cdot t^{-m}$ (favored by J. R. Anderson & Schooler, 1991; Rubin, 1982; Wixted & Ebbesen, 1991); the exponential in the square root of time, $y = b \cdot e^{-m\sqrt{t}}$ (favored by Wickelgren, 1972); and the hyperbola in the square root of time, $y = 1/(b + m \cdot \sqrt{t})$ (previously unconsidered). The data, however, whether considered as independent studies or a whole, could not distinguish among these four functions. That is, the data were not sufficiently precise to be able to reject any of the four functions.

Here we provide three continuous cued-recall and two continuous recognition data sets that have the precision to discriminate among alternative functions. A quantitative description of retention is useful for both practical and theoretical purposes. From a practical standpoint, the question of how long material or skills learned under specific conditions will be available is best answered quantitatively. Knowing the level of performance at specific times after learning is more useful than knowing that forgetting drops rapidly at first and then levels off. From a theoretical standpoint, the naive layperson might expect psychological theories of memory to make detailed quantitative claims about the course of forgetting. After all, a basic observation that makes memory a topic of interest is that people remember less with the passage of time. Psychologists once tried to consider retention in their theories (e.g., Luh, 1922; Wickelgren, 1972, 1974; Woodworth, 1938) and recently have been criticized from a variety of perspectives for no longer doing so (Brainerd, Reyna, Howe, & Kingma, 1990; Ratcliff, 1990; Slamecka & McElree, 1983).

The discrepancy between what an outsider might expect and what psychologists have done is a main reason that the approach used here is not committed to a single strong theoretical perspective on retention. Although sophisticated mathematical models of memory exist, they do not make strong predictions about the mathematical form of the retention function. Even Anderson's adaptive model, which favors the power function, is only committed to the claim that retention and recurrence of events in the environment are similar (J. R. Anderson, 1990; J. R. Anderson & Schooler, 1991). There is a circular problem that we hope to solve here. Because no adequate description of the empirical course of retention exists, models of memory cannot be expected to include it, and because no current model predicts a definite form for the retention function, there is no reason for experimenters to gather retention data to test the models. Here the description of the empirical course of retention is made both for its own sake and as a challenge and impetus to

David C. Rubin and Sean Hinton, Department of Experimental Psychology, Duke University; Amy Wenzel, Department of Psychology, University of Iowa.

We wish to thank R. B. Anderson for his comments and Michael Brown, Ben Cawizell, Michael Dombrowsky, Andy Lasswell, Anna Leja, Scott Martin, Michael Nelson, Michael Pass, Heather Rasmussen, Tammy Statler-Cowen, Erica Wagner, and Michelle Wille for their help in testing participants and scoring data, and Robert Jackson for help with programming.

Correspondence concerning this article should be addressed to David C. Rubin, Department of Experimental Psychology, Duke University, Durham, North Carolina 27708-0086. Electronic mail may be sent to david.rubin@duke.edu.

its inclusion, along with other memory phenomenon, in model building. We therefore include the values that most modeling efforts would require: amount remembered in terms of percentage correct and d'-type measures, errors in terms of recall intrusions and recognition false alarms, and reaction times. This reporting produces an atheoretical list, but such reporting is needed if our data are to be used in combination with other findings in developing and testing mathematical models. Where we can, we offer theoretical accounts for our data, but this is not our only goal in reporting our results.

The following six criteria are desirable to obtain data sets that can distinguish among different mathematical functions (Rubin & Wenzel, 1996):

1. There should be nine or more retention intervals, which would allow nonlinear iterative fits of functions with two or more parameters to be made and discriminated.
2. The study should have small standard errors for each of these retention intervals, which should be publicly reported (Loftus, 1993). That is, the values should be precise. Functions that do not remain within confidence intervals could be rejected.
3. As the four most successful functions from the Rubin and Wenzel survey are based on logarithmic (or logarithmic-like) scales, the new data set should have a large ratio of the most to least amount remembered and a large ratio of the longest to shortest retention intervals without obtaining indeterminate amount-recalled values of 0% or 100%.
4. In order for the time between presentation and testing to be unambiguous, each item should be presented only once.
5. The activity that fills the retention intervals should be constant throughout the experiment so that time is proportional to the amount of intervening material.
6. Ideal data would allow retention functions (with larger confidence intervals) to be plotted for individual participants to guard against attributing to the aggregate data a retention function that does not describe individuals.

There exist many retention data sets in the literature, and each one is superior to the ones we report for some specific purpose (see Rubin & Wenzel, 1996, for a brief review of 210 of them). Ours are not the best, the most general, or the most useful, but they are the most precise in that they report the smallest errors of measurement. Combined with the other five criteria, this facilitates our goal of deciding among competing functions, a goal which we could not meet with existing data sets.

One method that meets these requirements is a generalization of a continuous recognition procedure (Braun & Rubin, 1998; McBride & Dosher, 1997; Shepard & Teghtsoonian, 1961; Wickelgren, 1972, 1974) to also include cued recall. This method produced relatively smooth curves for Wickelgren, even when he used only 6 participants. Over the course of the experiment, each participant is tested at each retention interval, so that it is possible to obtain retention functions for each participant individually. For recognition, a word appears on a screen for several seconds and then is replaced by another word. Each word appears twice. The first time it is a new item; the second time it is an old item. The number of words between its two occurrences is the lag or retention interval. The participant presses a key to indicate whether the word is old or new. A similar procedure is used for cued recall. A pair of words appear on the screen and at some later time the first member of the pair appears alone as a cue. When the single cue word appears, the participant's task is to type the second member of the pair.

We adopted these procedures for two recognition and three recall conditions. One recognition condition was an old–new choice and the other was a remember–know–new choice so that we could study both the relation between these two commonly used tasks and the relation between remember and know judgments (Gardiner & Java, 1991). The three cued-recall conditions were designed to provide different levels of amount recalled to yield a family of recall curves. One condition involved the presentation of the original learning pair and cue in the same color, one condition involved the presentation of all stimuli in white, and one condition involved the presentation of the learning pair and cue in different colors. The condition in which the cue was in the same color as the original learning pair was intended to produce a higher level of recall than the condition in which all words appeared in the same color, which in turn was intended to produce a higher level of recall than the condition in which the color of the cue and learning pair varied randomly. Given the highly empirical nature of this study, we do not speculate on the nature of the retention function, but rather attempt to interpret our results in terms of existing theory in the Discussion once the data have been presented.

Method

Participants

All of the participants were undergraduates fulfilling a course requirement. Those in the three recall conditions were from the University of Iowa, and those in the two recognition conditions were from Duke University. The experimental tasks were long, difficult, and boring for most participants, and about 20% of the participants appeared to give up part way into the task, pressing either random response keys or not responding. To remove such participants, we set inclusion criteria. We set these after examining the distribution of responses of at least the first 100 participants in each experimental condition to ensure that criteria would exclude only people outside the normal distribution of those who attended to the task. For the recall conditions, participants had to be correct at least .60 of the time on Lags 0 and 1 combined (i.e., at retention intervals in which there were 0 and 1 intervening items). The average value for these lags for the remaining participants was .79 (SD = .09). For the old–new recognition, the participants had to have 25 or fewer no-response trials, which eliminated participants who stopped responding; a score of at least .50 on our [(hits − false alarms)/(1 − false alarms)] recognition measure for Lags 0 and 1 combined; and a false-alarm rate not greater than .80, which eliminated participants who adopted the strategy of always answering *old*. The means for the number of no responses, recognition measure, and false-alarm rate for the participants remaining in the old–new recognition condition were 2.98 (SD = 4.18), 0.84 (SD = 0.12), and 0.64 (SD = 0.07), respectively. For the remember–know recognition condition with remember and know responses combined and considered as old responses, these values were 4.99

(SD = 5.63), 0.84 (SD = 0.11), and 0.65 (SD = 0.08), respectively. The no-response criterion accounted for more than half of the eliminated participants in the recognition conditions.

To obtain 100 participants who met the inclusion criteria for the recall criterion, we had to test 114 participants in the matched-color-recall condition, 115 participants in the white-recall condition, and 122 participants in the random-color-recall condition. To obtain 100 participants in each recognition condition who met the inclusion criteria, we had to test 131 participants in the old–new–recognition condition (19 had more than 25 no responses, 4 had less than .50 corrected recognition on Lags 0 and 1, and 8 had false-alarm rates greater than .8) and 120 participants in the remember–know–recognition condition (10 had more than 25 no responses, 6 had less than .50 corrected recognition on Lags 0 and 1, and 4 had false-alarm rates greater than .8).

Materials

Depending on the condition, items were presented in one of eight colors on an otherwise dark computer screen: light gray, light blue, light green, light cyan, light red, light magenta, yellow, and white. In the three recall conditions, word pairs (i.e., paired associates) were presented at learning. At recall, the first member of the pair was used to cue the missing second member. The cue, or stimulus member of the paired associate, was a six-letter word, and the to-be-remembered item, or response member, was a four-letter word. All words were chosen from Kuçera and Francis (1967) to have frequencies between 10 and 100 per million. Proper names, plurals, words with apostrophes, and highly emotional words were excluded, producing a population of 520 target words. There were approximately twice as many possible six-letter cue words that met these same criteria. These words were sorted according to their last letter, and the first 520 words were included in the experimental task. The 270 four- and six-letter words needed for each session were selected randomly from among these words. For training trials, both words appeared in a row in the middle of the computer screen, with the cue word centered one third from the left edge, and the to-be-remembered word centered one third from the right edge. In test trials, the cue word appeared in the center of the screen. For the matched-color, white, and random-color conditions, the learning and test trials were in the same randomly selected color, white, and two colors randomly selected with replacement, respectively.

The recognition test used digit–letter–digit trigrams of the form used in Canadian postal codes. These nonpronounceable nonsense strings were used to reduce the high level of recognition we found in pilot work. The numbers for a string were chosen from the digits 1 to 9. Zero was not included because it could have been read as the letter "o," making the trigram easier to code. The letters were selected from among the orthographically similar set of uppercase letters K, V, W, Y, and Z. Single spaces were placed between the numbers and letter. Nine positions on the computer screen were defined by the 3 × 3 matrix of upper–middle–lower and left–center–right. The middle center position was used only for feedback. To increase the difficulty of the recognition conditions, we randomly selected with replacement the color and position of the trigrams for each of the two appearances of each trigram. Thus, color and position at learning could not be used as cues to recognition.

A single pseudorandom order frame of 200 trials was constructed to provide 9 learning and 9 test trials at each of 10 lags spread equally on a logarithmic scale. The lags were 0, 1, 2, 4, 7, 12, 21, 35, 59, and 99, where lag indicates the number of intervening learning or test trials. In addition to these 180 trials (9 repetitions at each of 10 lags at both learning and test), there were 20 filler trials used to fill in spaces in the order; half of these filler trials were unscored learning trials and half were unscored test trials. A place in this single pseudorandom order frame existed for each to-be-remembered item to appear once as a learning trial (as the second member of a pair for recall or alone for recognition) and once in a test trial (as an implied question following the first member of a pair for recall or alone for recognition). This pseudorandom order frame of 200 trials repeated twice in each recall session and three times in each recognition session, resulting in 400 recall and 600 recognition trials. This provided 18 scored recall or 27 scored recognition tests at each of 10 lags. There were also 30 filler trials at the beginning of each session that were not scored to avoid some of the primacy effects observed in pilot testing. Therefore, what we find here may hold only where interference has reached a high level. In retrospect, it would have been useful to make measurements throughout the session and examine these early trials separately in some analyses. However, in the overall analyses, the effects of these early trials, which occur before much interference is developed, would have been swamped by the large number of later trials and, by the nature of the task, would have had to have been only for the short lags. The final sequence contained 430 trials for recall conditions and 630 trials for recognition conditions. Six-letter cue words, four-letter to-be-remembered words, and recognition trigrams were assigned randomly without replacement to their places in these total sequences. Thus, each participant had a different random assignment of stimuli to places in the sequence, and each recall participant had different stimulus–response pairings.

Procedure

For both the learning and test trials in the recall conditions, the stimulus appeared on the computer screen for 5 s, followed by a 1-s blank screen. Test trials required participants to type in the words that had been paired previously with the cue. This sequence repeated 430 times for a total time of 43 min. The instructions were read to the participants simultaneously as the participants read them from the computer screen.

For the recognition conditions, each trial had the following sequence: 2-s trigram presentation, 1-s blank screen, 0.5-s feedback, 1-s intertrial interval for a total of 4.5-s per trial for 630 trials (or 47 min 15 s total session time). The feedback consisted of the word RIGHT in green or the word WRONG in red placed in the middle–center position. Thus, the feedback appeared in a color and location not otherwise used. Participants were instructed to press a particular key if they had seen a stimulus previously and to press a different key if they had not seen the stimulus previously. Tape labels were placed over these response keys, and a large label was placed at the top of the keyboard. The remember–know recognition condition differed from the old–new recognition condition in that participants were instructed to respond in one of three ways: consciously remembering seeing the stimulus (remember), recognizing a stimulus but having no conscious memory of experiencing it (know), and not recognizing a stimulus (new).

Scoring

All responses made while a stimulus was showing or in the 1-s blank following the stimulus were scored. Reaction time was calculated as the latency until the first response key was pressed. For the recall conditions, verbatim recall was accepted as well as any responses that were obvious typing or spelling mistakes. These included responses in which a single letter was missing, responses in which the letter that was wrong was adjacent to the correct letter on the keyboard, reversed order of adjacent letters in the correct spelling, common misspellings, and changes from singular to plural or in tense. For the recognition conditions, the first response was taken.

Results

Recall Experiments

Our main empirical goal was to obtain a more precise description of laboratory retention than exists so that we could disambiguate and limit possible mathematical descriptions of retention and investigate differences among experimental tasks. In doing this, we assembled a rich data set, one that will support more alternative interpretations than can be considered in one journal article by one set of investigators. We therefore present our results in as clear and theoretically neutral fashion as possible in tables and figures as well as providing our best theoretical understanding of the data. In all fits, the data were not transformed to produce linear regressions. Rather, $1 - (\Sigma(y_i - \hat{y}_i)^2/\Sigma(y_i - \bar{y})^2)$ was maximized and reported as variance accounted for, abbreviated in the tables as r^2, where the y_is are the observed values, the $\hat{y}_i$s are the values estimated by the function, and $\bar{y}$ is the mean of the observed values.

Figure 1 presents the retention functions for the three recall conditions combined. As is common in such plots, we provide standard errors as error bars unless otherwise noted. The values that went into this and most other figures are provided in the Appendix so that readers can entertain their own alternatives. In addition, the ebbs measure (Bahrick, 1965), which is the same as d' without the subtraction of the false-alarm term, is included. In Figure 1 and many of the figures that follow, few if any of the standard-error bars can be seen because most of them fall within the symbol used to mark the mean. Also note that our pilot work allowed us to produce a full range of probabilities of recall from near 0.0 to near 1.0. For Figure 1 only, we label the horizontal axis in terms of time, instead of lags, to provide the reader an idea of the time scales involved.

Figure 2 presents the retention functions for our three recall conditions considered separately. On the basis of the distinctiveness at cuing, we expected more recall in the matched color than the all-white condition and more recall in the all-white than the random-color condition. Although these predictions were met at each of the 10 lags, the differences were small. Thus, we combined all 300 participants into one group for all further analyses. Nonetheless, Figure 2 is a clear demonstration of three near replications producing similar retention functions.

Figure 3 presents the data from the 300 participants combined and regrouped into quintiles of approximately 60 participants each on the basis of their average probability of recall for all lags combined. The division points for the

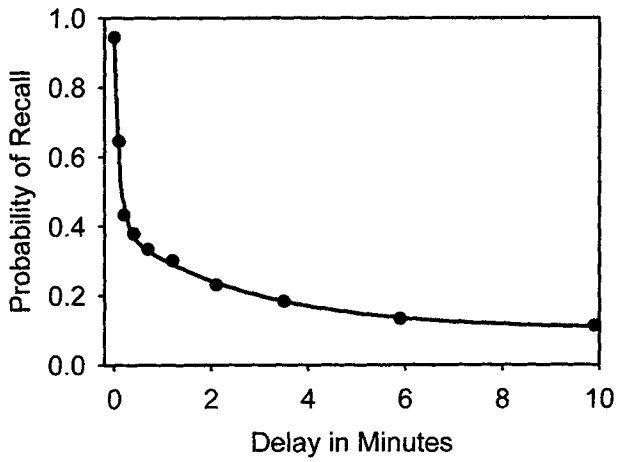

Figure 1. Probability of correctly recalling a word as a function of time since it was presented for all 300 recall condition participants. Error bars for standard error are included but are not visible because they are approximately .01 and thus are hidden by the plot points.

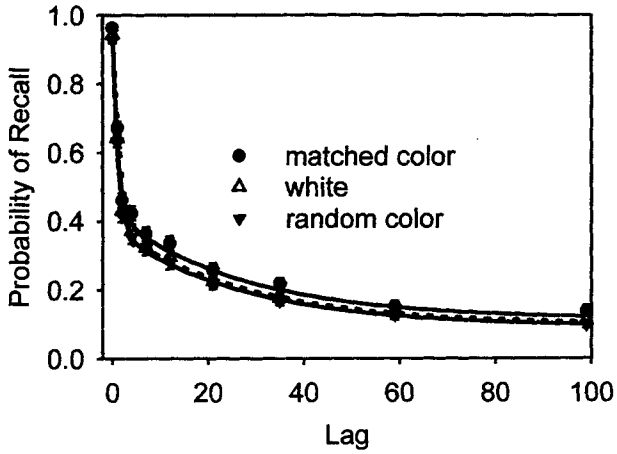

Figure 2. Probability of correctly recalling a word as a function of the number of items intervening since its presentation for conditions in which the color of the cue and to-be-remembered item were matched, randomly mixed, or all white. Error bars for standard error are included but are not always visible because they are often hidden by the plot points.

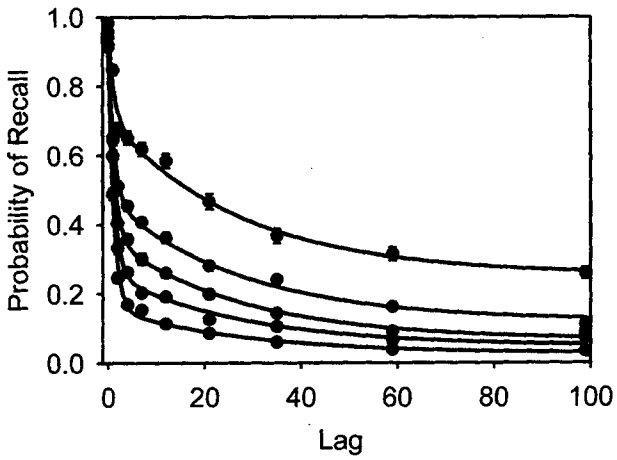

Figure 3. Probability of correctly recalling a word as a function of the number of items intervening since its presentation. The five plots are the data for 300 participants divided into quintiles on the basis of overall amount recalled. Error bars for standard error are included but are not always visible because they are often hidden by the plot points.

quintiles were average probability of recall equal to .465, .380, .315, and .265. Even though there are only 60 participants per group, grouping by overall level of recall keeps the standard errors small.

We began our attempt at fitting the data with the 105 two-parameter functions used by Rubin and Wenzel (1996). The best-fitting four functions from that study all fit the recall data from the combined 300 participants well, when the zero lag was removed to allow the logarithmic and power functions, which cannot fit lags of zero, to be included. The variance accounted for by the power is .973; by the hyperbola in the $\sqrt{t}$, .963; by the logarithmic, .942; and by exponential in the $\sqrt{t}$, .906. The five-parameter function sum of exponentials to be discussed shortly is a much better fit to the data as shown in Figure 1. When it is applied to the same nine data points with lag greater than zero, it accounts for .999 of the variance. A correction for the degrees of freedom was calculated for these five values by using the formula used by TableCurve (1994): $1 - (\mathrm{SSE/SSM}) \cdot ((n-1)/(n-p-1))$, where SSE is the sum of the residuals squared (i.e., $\Sigma(y_i - \hat{y}_i)^2$), SSM is the sum of squares about the mean (i.e., $\Sigma(y_i - \bar{y})^2$), n is the number of points plotted, and p is the number of parameters. With this correction, the five squared correlations just reported become .964, .951, .923, .875, and .998, respectively, and the differences between the two-parameter and five-parameter fits become even larger, suggesting that the additional three parameters are picking up more than just random variation. Because a sum of exponentials is a good fit, it follows from the work of Anderson (R. B. Anderson, 1996; R. B. Anderson & Tweney, 1997) that the power functions should be a good approximation, so the success of the power function is not surprising. General measures of goodness of fit aside, none of the four two-parameter functions were adequate because the more precise data collected here had systematic deviations for all four functions.

Figure 4 shows the fit of the best fitting of the four functions that were successful in Rubin and Wenzel (1996), the power function, to the five curves from Figure 3. Here the data are displayed on logarithmic axes so that the power function becomes a straight line. Also, for this figure and for one other like it for the recognition data, to be conservative (because we are arguing that the fit is poor instead of good), we display .05 level confidence intervals instead of standard errors. Five of the 5 lag 2 points are below their predicted curve (3 outside their .05 confidence interval), and 14 of the 15 lag 7, 12, and 21 points are above their predicted lines (5 outside their .05 confidence interval). For the groups with higher levels of recall, the data fall below the curves for the longest lags, whereas for the groups with lower levels of recall, the data fall above the curves for the longest lags. Overall, 13 of the 45 points are outside of the .05 confidence interval of their predicted fits, but this is not the main point; rather, it is that the deviations are systematic. A family of functions that bends to follow these systematic deviations would do better, as do the ones shown in Figure 3. In fact, for all 90 points plotted in Figures 1, 2, and 3, which contain all of the amount recalled data, only 1 point falls outside ±1 standard error, and it falls within the .05 confidence interval (i.e., within ±2 standard errors).

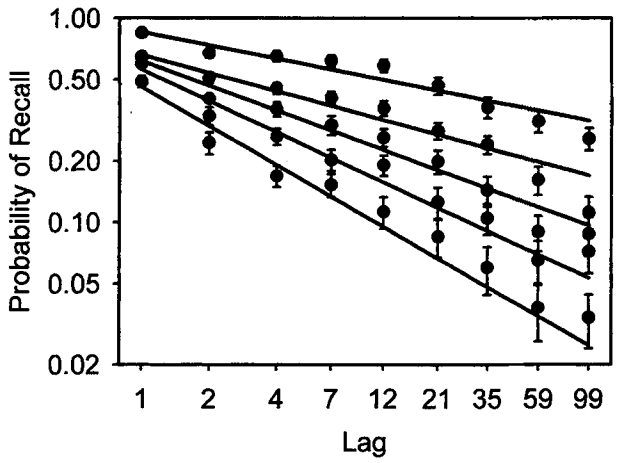

Figure 4. The same data as Figure 3 plotted on logarithmic scales and fit to power functions. Note the systematic deviations in that at Lag 2 all five points are below the lines fit; at Lags 7, 12, and 21, 14 of the 15 points are above the lines fit; and at Lags 35, 59, and 99 the points tend to converge to a middle value. For this figure, the error bars are $p < .05$ confidence intervals.

We tried many functions before settling on a sum of exponentials, including three- and four-parameter functions proposed in the literature and the four best-fitting two-parameter functions from Rubin and Wenzel (1996) extended by adding various third and fourth parameters. However, given the infinite number of possible three-, four-, and five-parameter functions, it is a near certainty that a function with five or fewer parameters will fit as well as or better than the one we found; we just could not find it after considerable time searching. We chose the function we did because (a) it best fit the complete set of data assembled, (b) it allows for a straightforward interpretation of the results in terms of existing theory, (c) it has more attractive mathematical properties than functions based on the logarithmic or power functions that are not well behaved at their limits, and (d) it makes use of the exponential that has a long history as a retention function both alone (Loftus, 1985; Peterson & Peterson, 1959; Wickelgren 1974) and in a sum of terms as used here (Daily, 1998; Daily & Boneau, 1995; Simon, 1966). More functions could be tried and a systematic comparison among them undertaken. All the data needed to do this are presented. However, we feel that a more fruitful path is to accept the function given as a good tentative mathematical description and to combine the data of this study with the other observations about memory that are usually considered in modeling.

The general version of the function we use here is $a_1e^{-t/T_1} + a_2e^{-t/T_2} + a_3e^{-t/T_3}$, though it is never used in that full form here. Rather, in all of its uses here, T_3 is ∞ and a_3 becomes the asymptote for long lags. In Figures 1, 2, and 3 and all that follow using a three-term function, unless specifically noted, T_1 and T_2 were fixed to be that of the best fitting line for all the recall data combined, making the curve

$a_1e^{t/1.15} + a_2e^{t/27.55} + a_3$. We fixed the T parameters and left the a parameters to vary, as it is customary in modeling to assume that the exponent describes a basic process, whereas the coefficient describes an initial level that is more likely to vary among individuals and conditions. Again a strong theoretical prediction could change such conventions. Thus, the eight data sets shown in Figures 2 and 3 were all fit to this three-parameter function. A limitation is that the two numeric constants were derived from the same data being fit. However, these values also work well for the recognition data to be described later.

We use the $-t/T$ form for the exponent instead of the $-bT$ form (i.e., e^{-bT}), which is more common in psychology, because the value of T is easier to interpret: Each time t increases by T, the value of $e^{-t/T}$ decreases to .37 of its original value. If a_1 were 1.0, then at $t = 0$, $a_1e^{-t/T}$ would be 1.0; at $t = T$, it would be .37; at $t = 2T$, .14 (i.e., $.37^2$); at $t = 3T$, .05 (i.e., $.37^3$); and at $t = 4T$, .02 (i.e., $.37^4$). The function we are using has three terms or processes. The first process has a T equal to 1.15, which we argue later is a reasonable estimate for working memory in our paradigm. By Lag 5, this process is adding little that can be measured to the retention function. The second process has a longer T equal to 27.55 and accounts for the shape for the middle of the curves. The third process, which has no discernible drop over lags up to 99 and so has a T equal to infinity, provides the asymptote.

The form of the equations initially fit to the data was the sum of three terms, which assumes that information is stored exclusively in one process corresponding to one term. However, if there were three independent processes and if information could be in each, then the coefficients of the processes would change. We would have to extend the standard $p(a) + p(b) - p(a) \cdot p(b)$, two independent process formula, to three processes. Thus, if an item could be in Process 1, 2, or 3 or more than one of them, the observed recall would be equal to the probability that it was in Process 1 or if not in Process 1 then in Process 2 or if not in Process 2 then in Process 3, or $p_1 + (1 - p_1) \cdot [p_2 + (1 - p_2) \cdot (p_3)]$. This mathematical description leaves the basic exponential components the same but changes their coefficients. Table 1 contains the values for the three free parameters using this independent-processes model for the data combined, for the three experimental conditions, for the data divided into quintiles, and for later recognition studies, along with the variance-accounted-for values and the increase in these values if T_1 and T_2 are allowed to vary. The reason for showing the fits to the five-, three-, and (where appropriate) two-parameter versions of the equation is to demonstrate that little predictive power is lost when the T_1, T_2, and (where appropriate) the a_1 coefficients are forced to take the values obtained by fitting the combined 300 participant recall data. That is, such losses are small enough to be caused by the five-parameter model changing to fit chance variation. The proportion-of-variance-accounted-for values in the table are always greater than .985, even with two or three free parameters, which supports the claim that the goodness of the fits is not caused by having five free parameters to adapt to any changes that occur in the points being fit, whether they are systematic or not.

These fits reveal an interesting property. For the recall data, the a_1 coefficients do not show a systematic monotonic change over quintile groupings. That is, the short time-constant process does not vary systematically as a function of the total amount recalled, whereas the other two parameters do. If the a_1 coefficients were all set to the value for the grouped data, then the resulting five groups would be described by equations with two parameters set from each group (a_2 and a_3) and three set from the groups combined (a_1, T_1, and T_2). This would result in an average drop in variance accounted for of only .0003 from the three-free-parameter values shown in Table 1 and of only .0013 from

Table 1
Fits of All Data Sets to the Function $a_1e^{-t/T_1} + a_2e^{-t/t_2} + a_3$

Data set	Parameter			r^2 with different no. of parameters			Change in r^2 with parameters	
	a_1	a_2	a_3	5	3	2	5 to 3	3 to 2
Cued recall								
Combined	.92	.32	.10	.9957				
Matched	.94	.35	.13	.9945	.9944	.9940	.0001	.0004
White	.92	.31	.10	.9960	.9960	.9960	.0000	.0000
Random	.89	.31	.09	.9957	.9957	.9953	.0000	.0004
1st fifth	.95	.61	.25	.9885	.9880	.9874	.0005	.0006
2nd fifth	.90	.40	.12	.9987	.9967	.9860	.0020	.0007
3rd fifth	.93	.32	.06	.9931	.9928	.9928	.0004	.0000
4th fifth	.92	.22	.05	.9929	.9911	.9911	.0020	.0000
5th fifth	.90	.14	.03	.9969	.9967	.9963	.0002	.0004
Intrusions	.10	.02	.00	.9957	.9956		.0001	
Recognition								
Old–new	.11	.80	.22	.9971	.9969		.0002	
Remember + know	.66	.63	.19	.9967	.9947		.0020	
Remember	.64	.44	.10	.9957	.9943		.0014	

Note. The a_1, a_2, and a_3 parameters shown are for the three-parameter fit with $T_1 = 1.15$ and $T_2 = 27.55$, except for intrusions for which $T_1 = 2.75$ and for old–new recognition for which $T_2 = 13.38$. The two-parameter fit for the recall fixes $a_1 = .92$.

the five-free-parameter solutions. In contrast, if either the a_2 or a_3 coefficients were set to the value for the grouped data, there would be an average drop in variance accounted for of .0376 and .0286, respectively, from the three-free-parameter values. Thus, the short time-constant term appears not to vary systematically, both in its a_1 and T_1 parameters.

In an attempt to further understand the processes underlying recall, we examined the distribution of the intrusions of response words from the list that were paired with an incorrect cue word. As there were only an average of 14.54 intrusions per participant, we pooled all intrusions from the three recall conditions. Because individual participants' data were not scored separately, the figure for these data, and this figure only, has no error bars. We took each of the lags used for the presentation of stimuli and summed the number of intrusions that occurred at that lag and any trial after the previously summed lag. This sum was divided by the number of trials included and then by 4,362, the total number of intrusions from all 300 participants, to give the probability of an intrusion coming from each of the lags. Thus, the number of intrusions at Lags 0, 1, and 2 were divided by 4,362; the number of intrusions at Trials 3 and 4 were summed and divided by 2 times 4,362 and assigned to Lag 4; the number of intrusions at Trials 5, 6, and 7 were summed and divided by 3 times 4,362 and assigned to Lag 7; and so forth till Lag 99. Because there were 430 trials in the experiment, intrusions could come from beyond trial 100, so we extended the logarithmic progression with "lags" at Trials 165, 275, and 429. This scheme was used because it produced intrusion data directly comparable with the probability-of-correct-recall data and because the logarithmic scale produced an approximately equal number of intrusions at each lag, resulting in an approximately equal accuracy at each lag.

Figure 5 presents the probability of intrusion as a function of lag for lags up to 100. The remaining three points not shown were 165, .0010; 275, .0005; and 429, .0002. The data were not fit well by the function used for the recall data unless the short time constant, T_1, was changed from 1.15 to its best-fitting value of 2.75, as was done for the curve in Figure 5. Moreover, as the values were very small at long lags, the asymptote, a_3, could be removed with little decrease in the fit. The greater value for T_1 occurred because there were fewer intrusions from the shortest lags, which could be because participants recognized these words as not the correct answer and did not give them as often as they come to mind. The 0.0 asymptote can be seen as what would occur as a natural extension from Figure 3, if a group of recall participants had as low an overall recall as there were intrusions. Aside from these minor differences, participants seemed to be retaining words to use as intrusions with the same distribution that they retained them for correct response.

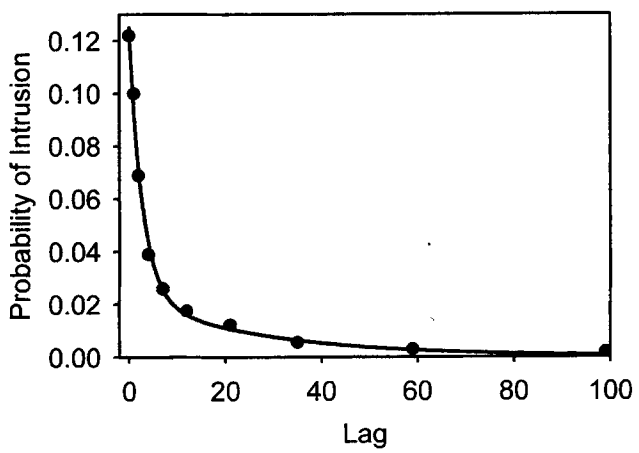

Figure 5. The probability of an intrusion as a function of the number of items intervening since its presentation.

Table 2
Source of Intrusions Measured in Number of Trials Before Test Item

Lag	*M*	*Mdn*	Mode	*n*	Q_1	Q_3	$Q_3 - Q_1$
0	14	3	2	90	2	7	5
1	21	3	1	454	1	7	6
2	36	13	1	407	4	36	32
4	38	8	2	493	3	27	24
7	53	15	3	464	5	62	57
12	42	9	2	550	3	37	34
21	45	16	2	450	5	47	42
35	55	25	2	460	4	76	72
59	59	21.5	1	496	4	79	75
99	62	20.5	2	498	3	92	89

Note. Q_1 and Q_3 are quartiles; n is the number of intrusions observed at each lag.

The participants also showed a sensitivity to how long ago the correct answers were presented. Table 2 lists the mean, mode, median, and quartiles for how many trials back the intrusions came as a function of the lag of the correct answer that was displaced by the intrusion. In addition to the tendency to recall intrusions from recent trials as shown in Figure 5, Table 2 shows a tendency to recall intrusions that were presented longer ago when the correct, but not given, answer occurred longer ago. Thus, the participants had an idea of where in the list the correct answer was, and they used that information in selecting the wrong answer.

Figure 6 shows the reaction times from the onset of the cue words until the first letter of the response was typed for correct and incorrect responses as a function of lag. Only the combined data are shown because there were no systematic differences among the three experimental conditions or quintiles. These data were also fit by a sum of two exponentials and an asymptote, but because there is little historical basis on which to argue for a specific function, and as testing of other functions with more than two parameters was not undertaken, this is just offered as a convenient fit. Thus, here and with other reaction time data, as opposed to amount remembered data, little can be made of the data points that are not within a standard error or two of their functions. What is clear is that errors take longer than correct responses and that reaction times rise sharply and then asymptote at well below the maximum 6-s value that is possible. Such an asymptote in reaction time is apparent even at much longer retention intervals (Reber, Alvarez, &

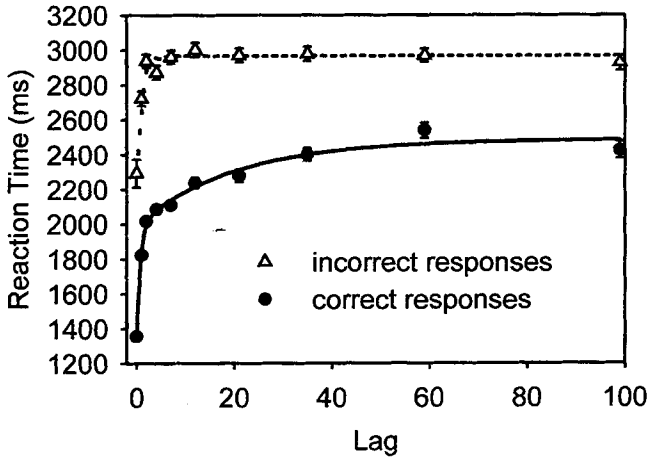

Figure 6. Reaction times for correct and incorrect responses as a function of the number of items intervening between presentation and test. Error bars for standard error are included.

Squire, 1997). The reaction time for Lag 0 is much shorter than the other reaction times.

Fits to Individual Participants' Recall Data

Anderson (R. B. Anderson, 1996; R. B. Anderson & Tweney, 1997) has shown, by a series of simulations, that averaging over individual power functions, exponentials, and truncated linear and logarithmic functions with reasonably distributed parameters can produce an aggregate power function. Their simulations do not show the deviations from a power function that our data show in Figure 4. Nonetheless, we cannot rule out the possibility that functions other than the sum of exponentials fit the individual participants' data and still averaged to the sum of exponentials. These other functions would have to be either the standard ones used by Anderson but with different distributions of parameters than they used or one of the infinite number of functions they did not use. Moreover, given that the individual participants' data are noisy, many functions should be indistinguishable from each other and the sum of exponentials. However, we can at least ensure that the sum of exponentials fits the individual participants' data and that the average value of the coefficients fit to the individual participants' data are consistent with the coefficients fit to the grouped data (Estes, 1956).

To investigate whether the function that fit the grouped data would hold for individual participants, we fit it to the individual data of all 300 participants who took part in the recall conditions. We chose these participants because (a) we could include the same two- and three-parameter versions for all 300 participants, and (b) we had information on the gender and American College Test (ACT) scores for 289 and 266 of the 300 participants, respectively, that could be correlated with the parameters of the fit. Note that in all fits reported in this section, the values of T_1 and T_2 were fixed at 1.15 and 27.55. In the three-parameter fits, the values of the a_1, a_2, and a_3 parameters were allowed to vary, and in the two-parameter fits the value of a_1 was fixed at .92 and a_2 and a_3 were allowed to vary. For the three-parameter fit, the average variance accounted for of the individual fits was .89 (SD = .09), and the average a_1, a_2, and a_3 parameters were .90 (SD = .17), .33 (SD = .20), and .10 (SD = .11) compared with values of .92, .32, and .10 for the grouped data. For the two-parameter fit, in which a_1 was fixed at .92, the average variance accounted for of the individual fits was .87 (SD = .11), and the average a_2 and a_3 parameters were .33 (SD = .20) and .10 (SD = .12). Because the individual participants' data were fit well by the same function that fit the grouped data, the results need not be seen as an artifact of averaging over participants; that is, the grouped data were representative of the individual data.

One could argue more strongly that the different components of the function represented different processes if the parameters of these components varied in systematic ways with individuals. To test this hypothesis, we correlated the parameters from individual participants with their gender and ACT score. Because the a_2 parameters of the two- and three-parameter fits correlated .95 with each other, and because the a_3 parameters of the two- and three-parameter fits correlated .97 with each other, we consider the two- and three-parameter fits as nearly identical and concentrated on the three-parameter fit because its a_1 parameter was also available. For the three-parameter fit, a_1 correlates with a_2 and with a_3 − .16 and −.16 ($ps < .01$), indicating that individuals who had higher a_1 values tended to have slightly lower a_2 and a_3 values. By contrast, a_2 and a_3 had a moderate positive correlation of .34 ($p < .0001$). ACT scores (M = 25.10, SD = 3.49) correlated with the a_1, a_2, and a_3 parameters −.05 (*ns*), .24, and .26 ($ps < .0001$). Men had lower a_1 parameters than did women (.86 [SE = .03] vs. .92 [SE = .01]), $t(287) = 2.82$, $p < .01$; did not differ on a_2 (.36 [SE = .03] vs. .32 [SE = .01]), $t(287) = 1.30$, $p = .19$; and had higher values on a_3 (.13 [SE = .01] vs. .09 [SE = .01]), $t(287) = 2.17$, $p < .05$. Thus, a measure of scholastic aptitude correlated with parameters associated with longer term memory but not working memory, and men and women differed on the parameters associated with the shortest and longest time constants. Although these results cannot offer strong evidence outside a predictive theoretical framework, they do suggest that the three parameters may be measuring different processes.

Recognition Experiments

We divided the number of correct remember and remember-plus-know responses at each lag by 27 (the number of presentations at each lag) to compute the probability of a hit. In scoring know responses, however, we assumed that people would make a know response only if they thought that the item occurred earlier but could not "remember" it. We therefore divided the number of correct know responses at each lag by 27 minus the number of remember responses at that lag. Because the false-alarm rate was high, we used (p(hit) − p(false alarms)) / (1 − p(false alarms)) as a measure of performance for each participant that was most directly comparable with the probability of a correct response used with the recall data. This measure ranges from

−1.0 to 1.0, but if there are at least the same number of hits as false alarms, it has the 0.0-to-1.0 range of a probability measure. Separate false-alarm figures were calculated for remember and for know responses in the remember–know recognition condition. We also calculated d' and present these data in the Appendix (see Tables A2 and A3).

The best-fitting four functions from the 105 two-parameter functions used by Rubin and Wenzel (1996) were also fit to the recognition data with the zero lag removed to allow the logarithmic and power functions, which cannot fit lags of zero, to be included. The variance accounted for the old–new recognition and the remember recognition judgments by the logarithmic are .967 and .982; by exponential in the $\sqrt{t}$, .963 and .972; by the hyperbola in the $\sqrt{t}$, .967 and .976; and by the power, .932 and .955. These eight values when corrected for degrees of freedom in the same manner as the recall values become .956, .976, .951, .962, .956, .968, .909, and .940, respectively. The five-parameter fit shown in Figure 7 for the old–new and remember judgments, when made to the nine nonzero lags, are .996 and .998, and when corrected for degrees of freedom become .989 and .994. Thus, as with recall, the numerical differences among the two- and five-parameter fits increase when the correction for degrees of freedom is applied, suggesting that the extra parameters are accounting for more than just random variation. Figure 7 presents the results of the old–new and remember–know recognition conditions with the remember and know responses combined into a single "old" response, allowing a direct comparison between the two recognition conditions. Figure 8 presents the same data fit to power functions. As with Figure 4, the axes are both logarithmic, so that the power function is a straight line. Also as with Figure 4, the error bars are .05 confidence interval. The confidence intervals of several points do not include their functions, but of more importance, again there is systematic variation that a more complex function can fit, as shown in Figure 7.

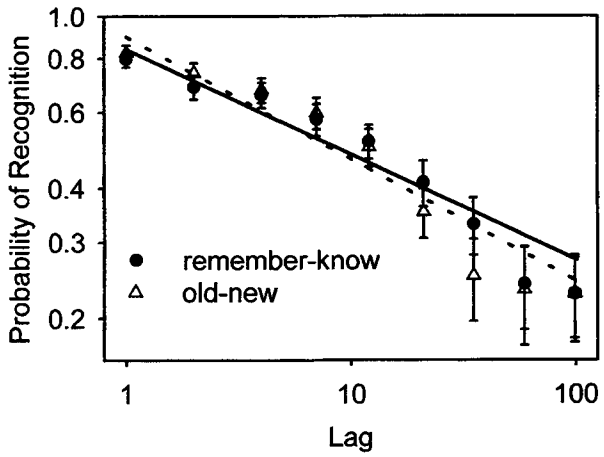

Figure 8. The same data as Figure 7 plotted on logarithmic scales and fit to power functions. For this figure, the error bars are $p < .05$ confidence intervals.

Returning to Figure 7, although the curves are similar in their overall level of performance, they differ in their shape. The old–new condition has more curvature, being well below the remember–know condition at Lags 21 and 35. Figure 9 reproduces the combined remember–plus–know response and gives its separate components. The remember–plus–know and remember curves shown in Figures 7 and 9 are fit to the $a_1e^{t/1.15} + a_2e^{t/27.55} + a_3$ function used for the recall data. The parameters are shown in Table 1 for the 1 (p(hit) − p(false alarms)) / (1 − p(false alarms)). As shown in Table 1, the variance-accounted-for values decreased little by dropping from five free parameters, which included T_1 and T_2, to only the three free a parameters.

For the know responses, a different set of parameters for T_1 and T_2 was needed. We started at Lag 1 instead of Lag 0 because no monotonically decreasing function could fit the

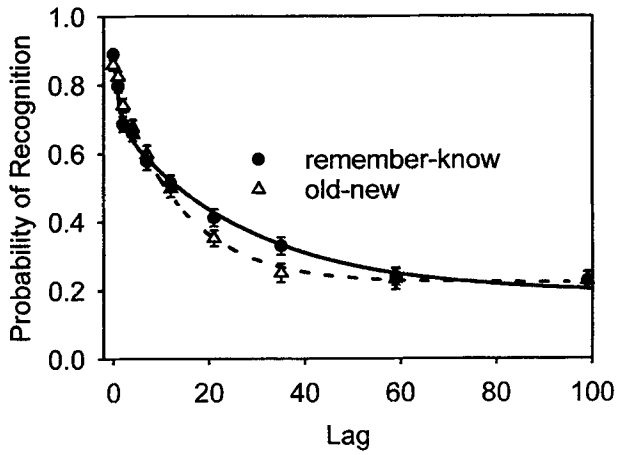

Figure 7. The probability of correctly recognizing a word as a function of the number of items intervening since its presentation for both an old versus new judgment and a remember versus know versus new judgment. The dependent measure is (hits − false alarms) / (1 − false alarms). Error bars for standard error are included.

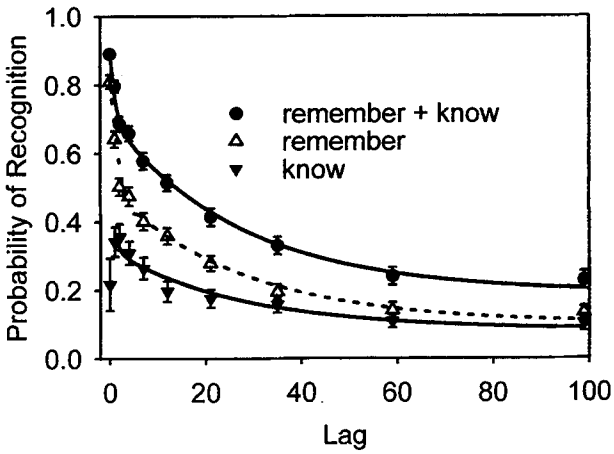

Figure 9. The probability of correctly recognizing a word as a function of the number of items intervening since its presentation as shown by a correct judgment that was remember or know, remember, or know. The dependent measure is (hits − false alarms) / (1 − false alarms). Error bars for standard error are included.

Lag 0 point. Participants were not likely to report that they knew but did not remember an item they had seen 1 s before with no intervening items. The best-fitting curve using the simpler exclusive process model was $.182e^{-t/7.76} + .116e^{-t/48.38} + .086$ with a variance accounted for of .9817 as compared with .9493 for the three-parameter version of the function that was fit to the recall data. Given that five free parameters were needed even with the first point eliminated, little confidence can be placed in this curve as any more than an empirical description. A nonmonotonic function is needed for more theoretical purposes, but the limited data do not support such a search here.

The old–new recognition data fit a different function. The equation shown in Figure 7 and noted in Table 1 has a T_2 value of 13.38. This T_2 value is different from the value in the equation that fit the recall data and the remember or remember–plus–know data of the remember–know condition. If that 27.55 T_2 value were used, the variance accounted for would drop by .0268. It appears that when people have to make both remember and know judgments, they do something somewhat different than when making only recognition judgments, something empirically more similar to recall. Perhaps they are searching for a context with which to make the remember–know distinction, a process not unlike searching for a context in recall. The a coefficients on the $T_1 = 1.15$ term is much less than in the remember–know condition. This difference may be due to the process described by that term being less or to the lower value of $T_2 = 13.38$ producing more curvature, which makes the T_1 term less necessary to describe the data.

We have used the (p(hit) − p(false alarms)) / (1 − p(false alarms)) corrected probability measure throughout our analyses because it is more directly comparable with the probability of recall measure than is d'. However, the basic conclusions do not change if we use d'. For the remember judgments, when all five parameters in our $a_1e^{-t/T_1} + a_2e^{-t/T_2} + a_3$ independent store equation are free to vary, the resulting equation is $.66e^{-t/1.00} + .51e^{-t/20.56} + .15$, $r^2 = .9983$. If we restrict T_1 and T_2 to the 1.15 and 27.55 values of the combined recall data as we did with the corrected probability measure, the variance accounted for drops by only .0013. For the old–new judgments, when all five parameters are free to vary the resulting equation is $.32e^{-t/1.85} + .78e^{-t/13.45} + .21$, $r^2 = .9979$. If we restrict T_1 and T_2 to the 1.15 and 13.38 values of the old–new corrected probability fit, the variance accounted for drops by only .0014. Thus, the probability and d' measure produce very similar results.

We examined the increase in interference by plotting the probability of a false alarm over the course of the two recognition conditions. The 630 trials of the experiment were divided into 21 bins of 30 trials each. The first bin of 30 trials contained the practice period. Although recognition data could not be scored for these trials, there were opportunities to indicate that an item occurred earlier when it did not. Figure 10 presents the false alarms for the incorrect yes responses from the old–new condition and the incorrect know and incorrect remember responses from the remember–know condition. The curves fit are two-parameter power functions, which resulted in slightly higher fits than three-free parameter exponentials and slight lower fits than five-free parameter exponentials, except for the Lag 0 point on the know data. Again there is little theoretical or historical guidance in selecting a function for these data, and the function is offered only as a convenient fit.

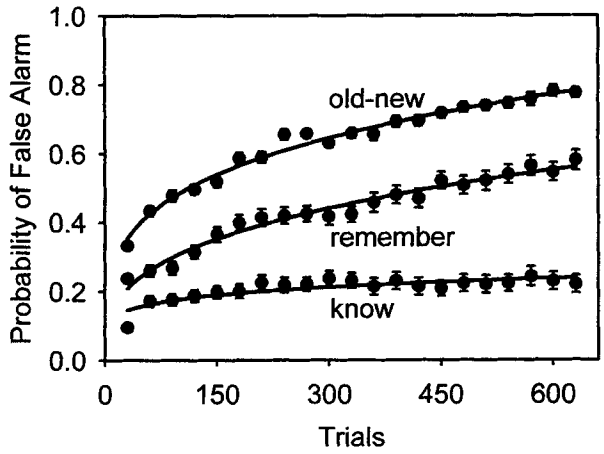

Figure 10. The false-alarm rate as a function of trials into the experiment, in blocks of 30 trials, old, remember, and know judgments. Error bars for standard error are included.

Reaction times for the old–new and remember–know conditions are shown in Figures 11 and 12, respectively. They are similar in form to those for the recall experiment and, like them, are fit to the five-parameter exponential equation for convenience. Note that the reaction times for the hits and misses for the old–new condition are nearly identical to those for the remember hits and the misses for the remember–know condition and that the know responses are longer than the misses. Having reaction times for a correct yes response longer than an incorrect no response (i.e., a miss) is surprising in that correct responses are typically shorter than errors. It offers support for the claim that know responses are made only after remember and new responses are rejected.

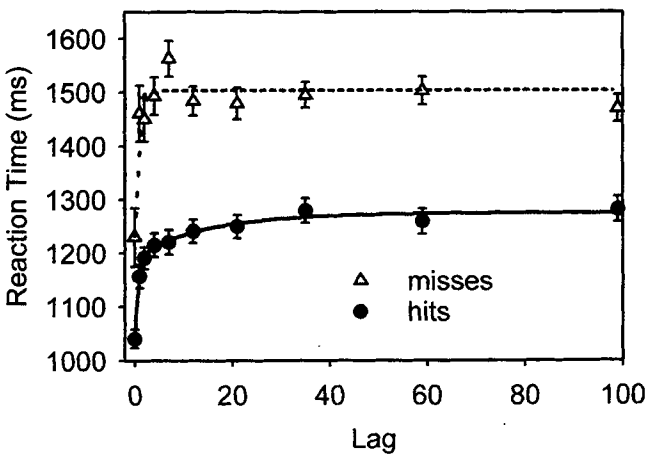

Figure 11. Reaction times for hits and misses in the old versus new recognition condition as a function of the number of items intervening between presentation and test. Error bars for standard error are included.

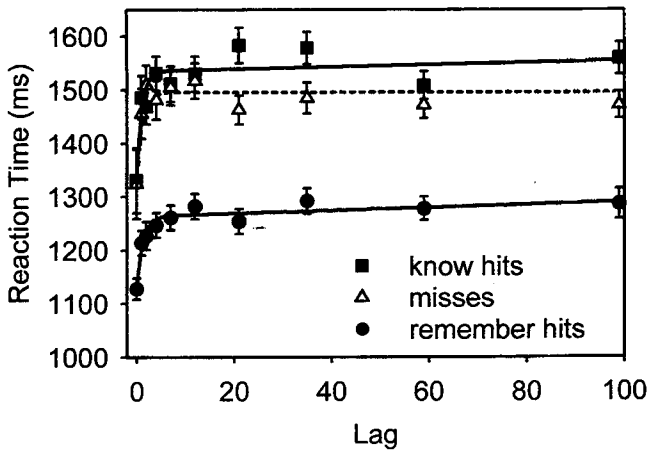

Figure 12. Reaction times for know hits, remember hits, and misses in the remember versus know versus new recognition condition as a function of the number of items intervening between presentation and test. Error bars for standard error are included.

Discussion

Overview

Here we provide the most precise data sets for retention yet to be obtained, achieved simply by testing more participants than have been tested previously in standard experimental procedures. We provide three continuous cued-recall and two continuous recognition data sets that have the precision to discriminate among alternative functions. We reject all the 105 two-parameter functions used by Rubin and Wenzel (1996) but fit the data well and in a theoretically satisfying way with a series of negative exponentials. That is, we provide the first data sets, outside autobiographical memory research, that are able to reject simple competing functions. In doing so, we raise some interesting theoretical possibilities about retention and argue for the value of quantitative analyses in memory research. We provide only five such data sets, leaving many questions about the range of generality of the findings and possible alternative theoretical explanations unanswered, but we do begin the precise study of retention.

Several conclusions can be drawn from the present study. From a methodological standpoint, it is clear that by increasing the precision of our results by testing more participants without any other methodological improvements over work done decades ago by Shepard and Teghtsoonian (1961) and Wickelgren (1972, 1974), it was possible to learn more about the nature of retention. At the least, these data provide memory researchers a precise quantitative description of a central phenomenon in their field—the time course of retention—for a few experimental conditions. Using this description, we were able to exclude a host of two-parameter functions about which there has been considerable debate in the literature and to substitute a more adequate mathematical description. In particular, the function $y = a_1e^{-t/1.15} + a_2e^{-t/27.55} + a_3$ is a good fit to continuous cued-recall and continuous remember–know recognition, where t is measured in number of intervening trials. This mathematical description holds for several data sets with a range of lags between 0 and 99, and it holds for both grouped and individual data. Thus, contrary to past speculation, the logarithmic, power, and many other functions that have been proposed are not adequate to describe this form of laboratory retention, which was always measured under conditions of high interference. Moreover, this finding implies that laboratory retention does not follow the same time course as autobiographical memory retention, which is best described as a power function, at^{-b} (Rubin, 1982; Rubin & Wenzel, 1996). In addition, we now have evidence that the fit to remember–know recognition judgments is the same as that of recall and that the fit to old–new recognition judgments is of the same form but requires a different time constant. The function for old–new recognition judgments is $y = a_1e^{-t/1.15} + a_2e^{-t/13.38} + a_3$. The data show that intrusions come from a distribution like that of recall with Lag 0 reduced. Finally, the reaction time data make sense in terms of earlier studies. They support the claim that the Lag 0 point (or the $a_1e^{-t/1}$ term) is caused by a different process than the later lags and that know judgments as opposed to remember judgments are very slow.

Generality to Implicit Memory Tasks

Our past work reviewing existing retention data sets (Rubin & Wenzel, 1996) suggests that the function proposed here with appropriate parameters may fit a wide variety of conditions because it makes similar predictions to the two-parameter functions that did. Although attempts at describing the limits and generalizability of our function remain, a recent article uses a similar enough procedure to let us investigate whether our functions would also fit implicit-memory data. McBride and Dosher (1997) used versions of the continuous recognition task to test cued recall, stem completion, and recognition separately under both shallow- and deep-processing conditions. They fit their data to a function that for each lag chose the larger of either a power function or a constant. Their shortest lag was 20, by which point our $T_1 = 1.15$ term would be too small to be of any use. We therefore tested a reduced form of our equation, $y = a_2e^{-t/T_2} + a_3$. When we used the more conventional sum rather than a choice of the maximum value, as McBride and Dosher did in their article, the exponential plus an asymptote was at least as adequate as the power plus an asymptote for all data, though we cannot claim either function to be superior with their data. For their Experiment 3 stem completion using semantic and graphemic processing and cued recall using semantic and graphemic processing and their Experiment 4 recognition using semantic and graphemic processing, our variance-accounted-for values were .95, .94, .97, .99, .86, and .96, respectively. Of more interest, our T_2 values were 34, 31, 47, 36, 39, and 45, respectively, not far from the values of 13 and 27 from our experiments. The larger values could be due to undetermined changes in experimental procedure or to participants adapting to the longer retention intervals (R. B. Anderson, Tweney, Rivardo, & Duncan, 1997). Consistent with McBride and Dosher's analysis, we found that forcing all T_2 values to be 38.5, their average value for the six conditions, reduced the variance

accounted for of the fits by an average of only .0028. Thus, there is some evidence that the function that best fits our continuous cued recall and recognition, with minor adjustment in parameters, will fit other similar continuous explicit tasks and also fit continuous implicit-memory tasks. Moreover, once within an experimental procedure, the values of the T parameters in McBride and Dosher's data as well as our own do not seem to change much with the particular task participants perform.

The Problem of Comparing Functions With Different Numbers of Free Parameters

One could ask whether our five-parameter function fits better than the two-parameter functions that have been used to model retention simply because it has more parameters. At one level the answer is yes. A two-parameter exponential does not provide a good fit to the data, but a series of them does. Adding exponential terms with free parameters clearly helps. But there is more to our use of a five-parameter function than this. The function we use to describe the retention data works well for other situations when we fix all but two of the parameters from the combined recall data. The remaining two free parameters vary in systematic and theoretically interpretable ways when we group the data on the basis of five levels of total amount recalled or when we investigate individual participants' data, and we need only minor theoretically interpretable changes in the three "fixed" parameters to account for recall intrusions and recognition. Moreover, we argue that the standard two-parameter functions fail, not because they account for less variance, but because the variance they fail to account for is systematic rather than random error variance. This is one reason to increase precision (Meehl, 1978).

Nonetheless, a problem remains in comparing our proposed function to the two-parameter functions that have been used previously that should be considered in more detail. Even when the five-parameter function is reduced to two free parameters for a subset of the recall data, it is reduced by using an aggregate of the same data that are being predicted. This lack of independence is the reason we do not in general adjust for the number of parameters when calculating variance-accounted-for figures, though when we did make such a correction to compare the five-parameter function with the standard two-parameter functions, such as the power, the numerical superiority of the five-parameter function increased. Thus it is clear that the better fit of functions of more than two parameters is not due to the extra parameters being used to fit only chance variation. The only true two-parameter versions of our five-parameter function occurred when we fit recall-intrusion and recognition data with parameters fixed by the recall data. This worked well for the old–new recognition data but required a change of one of the "fixed" parameters for the intrusion data and for the remember–know–new recognition data. This overall pattern of results gives us confidence that the five parameters are not being used to capture primarily random variation as opposed to reliable results.

Theoretical Interpretations

In general, our view has been that the way to advance our understanding of memory is not to try to derive a retention function or theory of memory based on just retention data, but to include quantitative descriptions of retention data in mathematical models that explain many memory phenomena (Rubin, 1982; Rubin & Wenzel, 1996). Where we make any theoretical claims consistent with the function, as we do next, we support them with data from errors or reaction times. In particular, the claim that the first term of our function, $a_1e^{-t/1.15}$, represents working memory depends on prior work on spaced practice (Braun & Rubin, 1998) and on the short reaction times of the Lag 0 recalls and recognitions, as if items from Lag 0 were immediately available rather than having to be retrieved. If a different function were adopted that did not have a term or parameter dependent on the first few lags, these other observations would still remain.

Thus, rather than starting with a theory, we began by trying to obtain a description worthy of theoretical description and extension. This means that the theoretical account to be offered here is post hoc with respect to the data collected, but it is consistent with other claims we have made earlier (Braun & Rubin, 1998; Rubin, 1995; Rubin & Wenzel, 1996). The general form of the function we fit is $y = a_1e^{-t/T_1} + a_2e^{-t/T_2} + a_3$, where T_1 is approximately 1 and T_2 is approximately 27 for most conditions but approximately half that value for old–new recognition. How can we interpret this function? We consider the first term (i.e., a_1e^{-t/T_1} with $T_1 = 1$) a description of working memory. At $t = 0$, it has a value of a_1; at $t = 1$, $.37a_1$; and at $t = 2$, $.14a_1$ (i.e., $e^0 = 1$, $e^{-1} = .37$, and $e^{-2} = .37^2$). Having working memory be reduced to .37 of its value with each intervening trial makes sense in terms of a 2-s storage in the phonological loop of working memory (Baddeley, 1997). This is because the numerous rehearsal, retrieval, and match processes being performed in the continuous recognition and recall tasks will in one trial nearly exhaust a 2-s working-memory phonological loop. In the recognition conditions, a trigram is shown in each trial, and the participant tries to match it with an earlier trigram. If the trigram is being shown for the first time, then many trigrams will be retrieved and compared in vain with the one being presented. If the trigram is being shown for the second time, many retrievals and comparisons are still likely. In the recall conditions, similar processing would occur for test trials, and active rehearsal would occur during the learning trials. Thus, in all cases, working memory will be taxed by even a single intervening trial. This argument depends on the first process being a working memory or temporary buffer rather than a more permanent store. Moreover, the clearing of working memory with one intervening item is consistent with the finding in the laboratory spacing literature that there are large differences between Lag 0 and Lag 1 (the spacing effect) but small, often nonsignificant, differences among lags longer than 0 (the lag effect; Braun & Rubin, 1998).

The reaction time data also support the claim that Lag 0, and possibly in the recall conditions Lag 1, depend heavily

on working memory (see Braun & Rubin, 1998). The reaction times from these lags are much shorter than those from all the later lags, which are all similar to each other. For the recall conditions the Lag 0 and Lag 1 reaction times are 942 and 476 ms faster than the mean of the longest 7 lags, which show little difference. For the old–new, remember, and know recognition conditions, the Lag 0 reaction times are 209, 144, and 212 ms faster than the mean of the longest 7 lags, which show little difference. This pattern would arise if the process of remembering a to-be-remembered item that had several intervening items spaced between its study and test trials was based on retrieval from long-term memory into working memory, whereas the remembering of items with no or sometimes with one intervening trial could be done directly from working memory without retrieval. In addition, the decreased values of know responses and recall intrusions from Lag 0 are consistent with direct access from a buffer rather than retrievals from long-term memory.

In the $y = a_1e^{-t/T_1} + a_2e^{-t/T_2} + a_3$ equation, the $a_2e^{-t/T_2} + a_3$ terms can either be considered as the description of one long-term memory process or as an intermediate and a long-term memory process. Although there has been considerable debate about dividing memory into a short-term and long-term store (Healy & McNamara, 1996), there has been less attention paid to dividing long-term memory into several stores of differing periods, despite some behavioral (Bahrick, 1984; Ericsson & Kintsch, 1995) and biological (Gibbs & Ng, 1977; Ng et al., 1991; Rosenzweig, Bennett, Colombo, Lee, & Serrano, 1993) suggestions. The biological data indicate that, in addition to a short-term memory, there are at least two distinct, pharmacologically dissociable, longer term memory systems. Although comparisons across both species and tasks are extremely speculative, the time spans of the longer term memories would not exclude the times as measured here. All that is intended at this point is to note that dividing long-term memory into components of differing duration is not without some possible support and cannot be excluded.

Moreover, we believe that the a_3 asymptote is not really a constant in time but represents a decline too small to detect in our experiment or even in experiments with considerably longer delays (McBride & Dosher, 1997). The possibility of a constant residue of recall until the experimental context changes, however, cannot be rejected. Whether we assume two kinds of processes for long-term memory or one depends on future work. If we were to opt for one process, however, it would have to be fit by the two terms, $a_2e^{-t/T_2} + a_3$, or another function that produces values very similar to it. In either case, the longer time constant implies that this process should be viewed as a more permanent store rather than a buffer that is continually being overwritten. It describes our participants' ability to retrieve items as a function of lag, both for items that are correct and items that are incorrect but present on the list that appear as intrusions.

We have no evidence that the information not remembered on a given trial is lost permanently, and there is much evidence from the history of the study of memory to indicate that at least some information can be recovered with the right cues. For instance, the same long-term components that describe correct recall also describe intrusions, and the level of recall increases when cuing with the same as opposed to a random color or no color. Thus we view our mathematical description as a retention function as opposed to a forgetting function. From this perspective, remembering involves selecting the correct investigator-requested item from among all other potential items in memory (Hunt & Smith, 1996; Rubin, 1995). Why does the ability to discriminate one item from among all others in memory decrease with trials? Interference from or confusion among the target and added items could be at fault. For shorter lags, the time (or number) of intervening trials that have passed is a good disambiguating cue (Baddeley, 1997, pp. 33–35). For longer lags it is not. Our analysis of the lags of intrusions as a function of the lag of the target shown in Table 2 supports this claim. There are increases in both the number of intervening trials since intrusions and the spread in the number of intervening trials of intrusions with increases in the lag of the target. If this view is correct, the added need to reinstate a context in recall or in making a remember–know–new judgment as opposed to just making an old–new judgment decreases the rate of loss of disambiguating information, as the value of the T_2 constant is twice as great for such conditions.

The theoretical account just offered is consistent with our data, but it is not the only account that could be given. We provide all of our key findings in the Appendix so that others can formulate their own accounts. We do this so comprehensive mathematical models of memory can now include the shape of the retention function and perhaps a quantitative description of the times taken to make a response, among the other properties of memory they attempt to explain.

References

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

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

Anderson, R. B. (1996, November). *The ubiquitous power law: A sign of underlying heterogeneity?* Poster presented at the 37th annual meeting of the Psychonomic Society, Chicago.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition, 25,* 724–730.

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

Baddeley, A. D. (1997). *Human memory: Theory and practice.* Philadelphia: Psychology Press.

Bahrick, H. P. (1965). The ebb of retention. *Psychological Review, 72,* 60–73.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113,* 1–27.

Brainerd, C. J., Reyna, V. F., Howe, M. L., & Kingma, J. (1990). The development of forgetting and reminiscence. *Monographs of the Society for Research in Child Development, 55*(3–4, Serial No. 222), 1–92.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: Evidence from once presented words. *Memory, 6,* 37–65.

Daily, L. Z. (1998). Multiple sources of priming in multi-trial recognition. *Dissertation Abstracts International, 59*(01), 433B. (University Microfilms No. AAG9822792)

Daily, L. Z., & Boneau, C. A. (1995). *Exponential decay in recognition memory: Some empirical results and a theoretical model.* Unpublished manuscript, George Mason University.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211–245.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140.

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition, 19,* 617–623.

Gibbs, M. E., & Ng, K. T. (1977). Psychobiology of memory: Towards a model of memory formation. *Biobehavioral Reviews, 1,* 113–136.

Healy, A. F., & McNamara, D. S. (1996). Verbal learning and memory: Does the modal model still work? *Annual Review of Psychology, 47,* 143–172.

Hunt, R. R., & Smith, R. E. (1996). Accessing the particular from the general: The power of distinctiveness in the context of organization. *Memory & Cognition, 24,* 217–225.

Kuçera, H., & Francis, W. H. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 397–409.

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition, 21,* 1–3.

Luh, C. W. (1922). The conditions of retention. *Psychological Monographs, 31*(Whole No. 142).

McBride, D. M., & Dosher, B. A. (1997). A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General, 126,* 371–392.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Ng, K. T., Gibbs, M. E., Crow, S. F., Sedman, G. L., Hua, F., Zhao, W., O'Dowd, B., Rickard, N., Gibbs, C. L., Sykova, E., Svoboda, J., & Jendelova, P. (1991). Molecular mechanisms of memory formation. *Molecular Neurobiology, 5,* 333–350.

Peterson, L. R., & Peterson, M. J. (1959). Short term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193–198.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

Reber, P. J., Alvarez, P., & Squire, L. R. (1997). Reaction time distributions across normal forgetting: Searching for the markers of memory consolidation. *Learning and Memory, 4,* 284–290.

Rosenzweig, M. R., Bennett, E. L., Colombo, P. J., Lee, D. W., & Serrano, P. A. (1993). Short-term, intermediate-term, and long-term memories. *Behavioural Brain Research, 57,* 193–198.

Rubin, D. C. (1982). On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior, 21,* 21–38.

Rubin, D. C. (1995). *Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes.* New York: Oxford University Press.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103,* 734–760.

Shepard, R. N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology, 62,* 302–309.

Simon, H. A. (1966). A note on Jost's Law and exponential forgetting. *Psychometrika, 31,* 505–506.

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.

TableCurve 2D [Computer software]. (1994). San Rafel, CA: Jandel Scientific.

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

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

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

Woodworth, R. S. (1938). *Experimental psychology.* New York: Henry Holt.

Appendix

Values Obtained in Experiments

Table A1
Probability of Recall

Lag	Matched	White	Random	All 3
0	.962 (.005)	.943 (.007)	.927 (.008)	.944 (.004)
1	.663 (.018)	.642 (.017)	.632 (.016)	.646 (.010)
2	.451 (.020)	.431 (.017)	.418 (.020)	.434 (.011)
4	.409 (.020)	.372 (.019)	.354 (.020)	.379 (.011)
7	.358 (.020)	.324 (.020)	.323 (.020)	.335 (.012)
12	.326 (.020)	.296 (.019)	.281 (.020)	.301 (.011)
21	.248 (.018)	.225 (.017)	.220 (.017)	.231 (.010)
35	.206 (.016)	.176 (.015)	.168 (.013)	.183 (.009)
59	.140 (.014)	.133 (.013)	.127 (.013)	.133 (.008)
99	.126 (.013)	.112 (.010)	.099 (.010)	.112 (.006)

Note. Standard errors are in parentheses.

Table A2
Probability of Recognition ((Hits − False Alarms)/(1 − False Alarms)) and False Alarms

Lag	Old–new	*r* + know	Remember	Know
0	.859 (.015)	.890 (.012)	.810 (.020)	.217 (.076)
1	.825 (.016)	.797 (.017)	.642 (.024)	.342 (.043)
2	.741 (.020)	.687 (.022)	.503 (.026)	.356 (.038)
4	.676 (.022)	.659 (.022)	.475 (.027)	.309 (.034)
7	.599 (.025)	.578 (.025)	.401 (.025)	.264 (.032)
12	.499 (.025)	.514 (.023)	.358 (.024)	.197 (.030)
21	.353 (.023)	.413 (.025)	.278 (.022)	.176 (.026)
35	.251 (.027)	.330 (.025)	.195 (.022)	.158 (.024)
59	.233 (.030)	.240 (.026)	.141 (.023)	.109 (.019)
99	.228 (.024)	.228 (.026)	.134 (.018)	.104 (.024)
FA	.643 (.007)	.649 (.008)	.436 (.019)	.213 (.018)

Note. Standard errors are in parentheses. FA = false alarms.

Table A3
Values for Ebbs for Recall and d' for Recognition

Lag	Match	White	Random	Old–new	Remember	Know
0	1.777	1.583	1.455	1.312	1.328	0.394
1	0.420	0.363	0.338	1.170	0.905	0.518
2	−0.123	−0.174	−0.206	0.973	0.679	0.504
4	−0.229	−0.326	−0.373	0.836	0.609	0.471
7	−0.364	−0.455	−0.458	0.691	0.507	0.377
12	−0.451	−0.536	−0.580	0.558	0.443	0.258
21	−0.681	−0.755	−0.772	0.374	0.344	0.255
35	−0.822	−0.932	−0.963	0.264	0.233	0.228
59	−1.080	−1.113	−1.140	0.229	0.176	0.149
99	−1.145	−1.218	−1.288	0.223	0.158	0.146

Table A4
Reaction Times in Milliseconds for Correct Responses and False Alarms (FA)

	Cued recall				Recognition		
Lag	Matched	White	Random	All 3	Old–new	Remember	Know
0	1,335 (21)	1,344 (19)	1,390 (23)	1,356 (12)	1,041 (17)	1,128 (20)	1,331 (61)
1	1,800 (36)	1,815 (36)	1,850 (31)	1,822 (20)	1,157 (22)	1,214 (23)	1,486 (40)
2	2,029 (47)	1,924 (34)	2,097 (47)	2,017 (25)	1,191 (20)	1,227 (26)	1,469 (33)
4	2,138 (52)	2,016 (40)	2,104 (47)	2,086 (27)	1,215 (22)	1,247 (23)	1,531 (32)
7	2,117 (51)	2,116 (50)	2,100 (45)	2,111 (28)	1,221 (23)	1,261 (23)	1,511 (33)
12	2,278 (55)	2,220 (47)	2,215 (55)	2,238 (30)	1,241 (22)	1,282 (24)	1,530 (33)
21	2,300 (55)	2,264 (67)	2,272 (59)	2,279 (35)	1,250 (22)	1,254 (23)	1,583 (33)
35	2,362 (62)	2,385 (57)	2,461 (72)	2,402 (37)	1,280 (23)	1,292 (24)	1,578 (30)
59	2,502 (77)	2,536 (76)	2,584 (76)	2,540 (44)	1,259 (24)	1,278 (22)	1,508 (26)
99	2,408 (83)	2,474 (84)	2,394 (84)	2,427 (48)	1,282 (24)	1,287 (28)	1,559 (30)
FA					1,294 (23)	1,314 (21)	1,579 (28)

Table A5
Reaction Times in Milliseconds for Recall Error and Recognition Misses

	Cued recall				Recognition	
Lag	Matched	White	Random	All 3	Old–new	r + know
0	2,345 (183)	2,114 (105)	2,421 (135)	2,292 (81)	1,230 (55)	1,324 (65)
1	2,792 (66)	2,666 (79)	2,711 (70)	2,722 (41)	1,461 (52)	1,456 (47)
2	3,060 (73)	2,886 (60)	2,871 (62)	2,938 (38)	1,450 (41)	1,509 (37)
4	2,925 (69)	2,849 (72)	2,846 (69)	2,872 (40)	1,493 (35)	1,481 (36)
7	2,980 (61)	2,970 (66)	2,931 (73)	2,960 (38)	1,563 (33)	1,505 (33)
12	3,055 (70)	2,985 (81)	2,962 (62)	3,001 (41)	1,484 (27)	1,517 (33)
21	3,078 (65)	2,945 (68)	2,888 (67)	2,970 (39)	1,479 (29)	1,463 (27)
35	2,989 (73)	3,053 (70)	2,890 (61)	2,978 (40)	1,495 (24)	1,485 (29)
59	3,022 (65)	2,926 (68)	2,958 (65)	2,969 (38)	1,503 (26)	1,472 (25)
99	2,962 (71)	2,769 (69)	3,047 (74)	2,927 (42)	1,470 (25)	1,472 (25)

Note. Standard errors are in parentheses.

Received April 9, 1998
Revision received December 1, 1998
Accepted December 14, 1998 ■