Memory & Cognition
2005, 33 (5), 793-810

Remembering the news: Modeling retention data from a study with 14,000 participants

M. MEETER
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
and Universiteit van Amsterdam, Amsterdam, The Netherlands

J. M. J. MURRE
Universiteit van Amsterdam, Amsterdam, The Netherlands
and Universiteit Maastricht, Maastricht, The Netherlands

and

S. M. J. JANSSEN
Universiteit van Amsterdam, Amsterdam, The Netherlands

A retention study is presented in which participants answered questions about news events, with a retention interval that varied within participants between 1 day and 2 years. The study involved more than 14,000 participants and around 500,000 data points. The data were analyzed separately for participants who answered questions in Dutch or in English, providing an opportunity for replication. We fitted models of varying complexity to the data in order to test several hypotheses concerning retention. Evidence for an asymptote in retention was found in only one data set, and participants with greater media exposure displayed a higher degree of learning but no difference in forgetting. Thus, forgetting was independent of initial learning. Older adults were found to have forgetting curves similar to those of younger adults.

Starting with Ebbinghaus (1885), memory researchers have attempted to find mathematical functions that might describe the shape of the retention curve. Some of the proposed functions were purely descriptive (e.g., the exponential, power, or logarithmic curves), whereas others were based on more or less detailed models of memory (Chessa & Murre, 2002; Wickelgren, 1974; Wickens, 1999). Both types of functions have been successfully fitted to large numbers of retention curves.

Nevertheless, many questions surrounding the retention curve are still unanswered. For example, it is still unclear whether the rate of forgetting is or is not independent of initial learning (Bogartz, 1990; Loftus, 1985; Slamecka & McElree, 1983) or whether older adults forget faster than do younger adults (Brainerd, Reyna, Howe, & Kingma, 1990; Cohen, Stanhope, & Conway, 1992; Wheeler, 2000). One reason for the long life of these controversies is disagreement about what would constitute a proper answer to the questions; in particular, how to measure and compare rates of forgetting has been hotly debated. Some researchers have suggested that a rate of forgetting is only meaningfully measured within a model of retention (Bogartz, 1990; Rubin & Wenzel, 1996), and in that case, whether or not different conditions exhibit the same level of forgetting becomes a question of whether a decline parameter has the same value when the model is fitted to those conditions.

Unfortunately, whether two conditions yield the same decline parameter value is not independent of the forgetting function used: Conclusions about parameter values are often bound by the model in which the parameters in question play a role (Rubin & Wenzel, 1996). Therefore, an ideal study of (e.g.) the dependence of forgetting on initial learning would fit several models to the data, so that dependence or independence could be corroborated according to different models.

In this article, we will present a study in which retention for news events was tested for around 14,000 participants and 500,000 data points. The participants were Internet volunteers who could log into a Web site after giving relevant personal details, and take a test in which they answered questions about news events, such as *What was the name of the American country singer who died on September 12, 2003?* (Question 1,430). Our primary goal while developing the Web site was to create a new retrograde amnesia test by submitting news questions to Web controls to test the appropriateness of the questions for inclusion in the test itself (Meeter, Murre, & Janssen, 2005). However, the control data are interesting in their own right and can be used to study retention and forgetting. For each participant, 30 or 40 questions were sampled concerning news events that had occurred at different moments in time. In this way, re-

This study was supported by a PIONIER Grant from the Dutch National Science Foundation (NWO) to the second author. We thank Jeroen Raaijmakers for stimulating discussions. Correspondence concerning this article should be addressed to M. Meeter, Department of Cognitive Psychology, Vrije Universiteit Amsterdam, Vd Boechorstraat 1, 1081 BT Amsterdam, The Netherlands (e-mail: m.meeter@psy.vu.nl).

tention was measured at intervals ranging from a single day to up to 2 years.

The sheer size of the data set in the study allowed models to be fitted to the data and rejected with a high degree of precision. Several retention functions were fitted to the data, and two of them are of particular interest: the *memory chain model* (MCM) and what we will here refer to as the *extended Weibull model*. Mathematical details of both models are given in the Appendix.

The recently proposed MCM has been fitted successfully to many forgetting data sets (Chessa & Murre, 2002). It assumes that underlying memory strength may be modeled as a number of points in memory, recovery of which would lead to the correct output. These points may be copies of the memory or stored details that, if remembered, could trigger retrieval of the correct answer. Such a retrieval attempt is conceived of as putting a window over memory: If the window contains one or more points, the memory is counted as retrieved. The points can disappear, modeling forgetting, but they can also be copied into more permanent stores, such as from working memory into long-term memory, or within long-term memory from a hippocampal store to a more permanent neocortical store.

The basic form of the second model, which Rubin and Wenzel (1996) call the *Rubin–Wenzel–Wickelgren–Weibull–Williams–Watts exponential power law*, has a long history in memory psychology and also fits retention curves rather well (Rubin & Wenzel, 1996). The formulation that Wickens (1998, 1999) gave it has an important advantage over other formulations of the same function, in that it parameterizes several features of retention functions. For example, memory decline may not be constant over all retention intervals: Forgetting is usually fast immediately after acquisition, then gradually decelerates at greater latencies (technically, the hazard function derived from forgetting curves is usually decreasing). Therefore, one parameter in this retention function sets the balance between early and late forgetting. Such parameterization has the advantage that questions about retention (e.g., whether forgetting is slower at greater latencies than at shorter ones) become a matter of model fits: specifically, of whether or not the model fits better with a certain parameter set to its optimal value rather than restricted to a default value (e.g., to 1).

A second aspect of retention that has been parameterized by Wickens (1998, 1999) is whether performance is perfect if the retention interval is extrapolated back to $t = 0$, and yet a third theoretically interesting aspect of retention that is parameterized is the final asymptote. Decline in retention may continue until performance is at last equal to 0, or performance may asymptote at a level above 0 or chance. Such an above-0 asymptote is what one would expect if some form of permastore exists in memory (see, e.g., Bahrick, 1984; Bahrick, Bahrick, & Wittlinger, 1975; Rubin, Hinton, & Wenzel, 1999). Along with the basic decline parameter, these three components lead to a four-parameter function (see Table 1 and the Appendix).

Other retention functions often do not parameterize these separate components, but they do stipulate whether or not recall starts at 1, whether there is a final asymptote, or what the balance is between early and late forgetting. With a logarithmic retention function, for example, immediate performance is equal to the value of a parameter that is not typically equal to 1, there is no positive asymptote, and forgetting is steeper at early stages of retention than at later stages.

The memory chain model does not naturally lend itself to perfect recall at the onset of retention. The balance of early and late forgetting is determined in the interplay between its parameters, but if measured with recall probability, it typically includes a short period with slow forgetting, then a steeper decline, followed by less pronounced forgetting (Chessa & Murre, 2002). Whether or not the forgetting curve exhibits an asymptote in forgetting is parameterized within the model. In one variant, the model assumes consolidation from a first store to a second, more permanent one. If forgetting from the second store is set at 0, the MCM exhibits an asymptote. This allows a model with an asymptote to be tested against one without.

Another aspect of forgetting that has generated interest is whether or not forgetting is equal for different conditions. In the models, this can be translated into the question of whether decline parameters are shared by conditions such as different levels of initial learning, older versus younger adults, or recall versus recognition.

The aspects discussed above can be translated into three questions about retention, which, as already out-

Table 1
Functions Used in This Study to Fit the Retention Data

Amended power:	$y = b + (1 - b)\mu(t + 1)^{a}$
Extended Weibull:	$y = b + (1 - b)\mu^{-(at/d)^{d}}$
Memory chain model (MCM):	$y = 1 - \exp\left\{-\mu_1\left[e^{-a_1 t} + \frac{\mu_2}{a_2 - a_1}\left(e^{-a_2 t} - e^{-a_1 t}\right)\right]\right\}$

Note—For explanations of the formulas, see the Appendix. For easy comparison, parameters have all been given labels corresponding to equivalent parameters in other models. Here, a refers to parameters setting the speed of decay, b is an asymptote parameter, d sets the balance between early and later forgetting, and μ refers to parameters setting the strength of initial performance and, in the MCM, of consolidation.

lined, can in turn be translated into hypotheses about parameter values. This allows the testing of different hypotheses by comparing models with more or fewer restrictions. We used two families of models with very different mathematical structures, the memory chain model (Chessa & Murre, 2002) and the extended Weibull function (Wickens, 1998, 1999), so as to untie our hypotheses from the specific model used. The Appendix provides mathematical formulations of the models and discusses which parameter can be identified with which hypothesis. An amended power function (one starting at a retrieval probability of 1 instead of at infinity) was also fitted to our data and was also used to investigate the issue of an asymptote in forgetting.

We tested two hypotheses related to the issues discussed above (which model variant was used to test which hypothesis is described in the Appendix).

1. That there is an asymptote in forgetting. In the extended Weibull function and the amended power curve, testing this hypothesis takes the form of testing a model with an asymptote parameter against one without such a parameter. In the memory chain model, a model with consolidation to a second store but without forgetting from this store must be tested against both a model without consolidation and one with both consolidation to and forgetting from Store 2.

2. That different conditions can be fitted with certain shared parameter values. In particular, we tested whether decline parameters in both models remained constant for participant groupings with respect to age and initial learning and for the two question formats (recall and recognition).

We tested these hypotheses using a Web site containing both an international news test in the English language and a test in Dutch aimed at the Netherlands. Since the questions and samples were different for each test, the two data sets generated will be discussed separately as Experiment 1 (Dutch test) and Experiment 2 (international test). We fit the functions discussed above to both data sets, in order to see how well models restricted by our hypotheses would fit the data. In Experiments 1 and 2, data gathered up to February 21, 2003, were analyzed, and for Experiment 3, we used data gathered from that date until February 14, 2004.

EXPERIMENT 1

Method

The Internet news test, which we call the *Daily News Memory Test* (DNMT), was part of a larger Web site about memory aimed at the general public. On our home page (www.memory.uva.nl), participants were enticed to "test their memory" with the DNMT. The site with the test went public on November 1, 2000, and is still operational. For Experiment 1, however, we restricted our analyses to data gathered up to February 21, 2003. By then, the Dutch version of the test had been completed 8,244 times. The data sets discussed here and for the other two experiments in this study can be found at www.neuromod.org/datasets.

Creation of the Questions

Questions for the test were created according to a tight script. Each working day, one of us (S.J.) searched through a large daily newspaper and the Web sites for television newscasts. Topics that had front-page attention in both media were deemed suitable for questions. Moreover, only those stories that described datable events were selected. A headline about this topic was then transformed into a question by taking one of the roles out and replacing it with an interrogative clause. This substitution guaranteed that each question had a simple, determinate, and unambiguous answer. Care was taken not to formulate questions in such a way that later news would include the answer (e.g., not *Who won the 2000 presidential election in the United States?*, since the answer to this question would be contained in every news bulletin about Mr. Bush).

For the recognition version, three lures were created by freely associating on the basis of either the answer or other parts of the headline. In the case of the question mentioned in the introduction (*What was the name of the American country singer who died on September 12, 2003?*), these lures were *Willie Nelson*, *Waylon Jennings*, and *Kris Kristofferson*. The participants were presented with the question and the four possible answers in random order and were required to select an alternative before they could proceed to the next question. The recognition version thus used a four-alternative forced choice (4AFC) format.

Those questions for which the answer was not only unambiguous, but short as well, were also prepared in open form. The formulation of the question was the same as in the 4AFC format, but participants were then presented with a text field in which they could type an answer. Scoring of these answers occurred automatically by matching the participant's answer against a word or partial word indicative of the correct answer. Spelling mistakes were neutralized by also matching on variants of the correct spelling of the answer. For the example question given above, not only the answer *Johnny Cash* was counted as correct, but also any answer that contained the string *John* or *Cash*.

On some days, no news event occurred of sufficient prominence, but on others more than one question could be formulated. In all, 1,006 Dutch questions have been created in the 4 years analyzed in this study.

Questions that proved too difficult were uninformative for the purpose of studying retention. We therefore checked both after 30 and 60 days whether participants surpassed chance performance on the question in the 4AFC format. When performance was below chance (i.e., when less than 25% of the participants confronted with the question answered it correctly), the question was removed from the item list. This occurred for 92 of the Dutch items (not included in the total of 1,006 questions), and the data for these items have not been included in the results presented here.

Design of the Test

The questions were entered into a database, together with the correct answer, the alternative spellings of the correct answer, the lures, and the date on which the event occurred. Tests were generated automatically by scripts that selected questions from this database for presentation to a participant. Answers provided by the participants were stored in the same database.

At participants' first use of the site, they had to register and answer some basic biographical questions. Our reasons for eliciting this information were stated to the participants, and privacy was guaranteed. If participants had already taken the test, they could log in with a user name and password for retesting. (Around 34% of the participants took the test more than once. It was possible for a participant to redo the test without logging in but by registering instead under a different name, but because registering took several minutes and logging in just a few seconds, it is unlikely that this occurred often.)

Once participants were either registered or logged in, they read a short set of instructions and were then presented with the questions. When the site went online, it contained only the Dutch test, which then consisted of 40 4AFC questions. From June 16, 2001, the format changed to 10 questions in open format and 20 in closed 4AFC format.

The questions were not chosen entirely randomly; for several reasons, they were sampled more from recent than from remote time periods. The questions were sampled without replacement one at a time at the moment that a participant submitted an answer for the previous question to the site's database. Thirty percent of these new questions were sampled from those created in the last 30 days, another 30% from approximately a month prior to that (from 31–60 days before the test day), and the remaining 40% from before that time. Each test was thus a stratified sample with respect to retention intervals.

Participants

Participants could come into contact with the DNMT in one of three ways: First, they could inadvertently encounter the Web site while surfing the Internet. Academic psychologists, for example, may have found the site through links on research sites. Second, they could have found it via a search engine. The site was indexed by several robots and regularly turned up as an entry on searches for *memory* or *memory improvement*, since the Internet site of which the DNMT was a part also contained a short memory improvement course. The third, and perhaps most common, way was through word of mouth. We encouraged this by giving our participants the option of sending friends an e-mail with their score on the test and a challenge to beat this score.

By February 21, 2003, the Dutch version of the DNMT had been completed 8,244 times by 4,239 participants. Incomplete tests were discarded, as were the data from participants who registered with improbable dates of birth (e.g., implying an age of 100 or less than 5 years old). In line with what has been reported about the Internet population (e.g., by Buchanan & Smith, 1999), the participant group was disproportionately well educated and tilted toward younger adults (see Figure 1 for a comparison with the general Dutch population). Sixty percent of the participants were male, 40% female.

Results

Performance and Predictors

In general, the DNMT proved to be moderately difficult. The average score was 42% correct for the questions in open format and 65% correct for the 4AFC questions. Performance dropped with the age of the question; for open questions, the proportion correct at the greatest interval between formulation of the question and test date was about one half of initial performance, and for 4AFC questions, this figure rose to around two thirds of initial performance.

Participants' average scores were analyzed as a function of the information that they gave about themselves in the process of registering: age, sex, highest educational degree obtained, and their self-reported average exposure to newspapers and TV news. In Table 2, normalized and rescaled raw regression weights are given. The amount of newspaper reading was the best predictor, followed by gender (with an advantage for males), education, TV news consumption, and age (with an advantage for older adults, undoing the effect of a generally lower level of educational attainment for this subgroup).

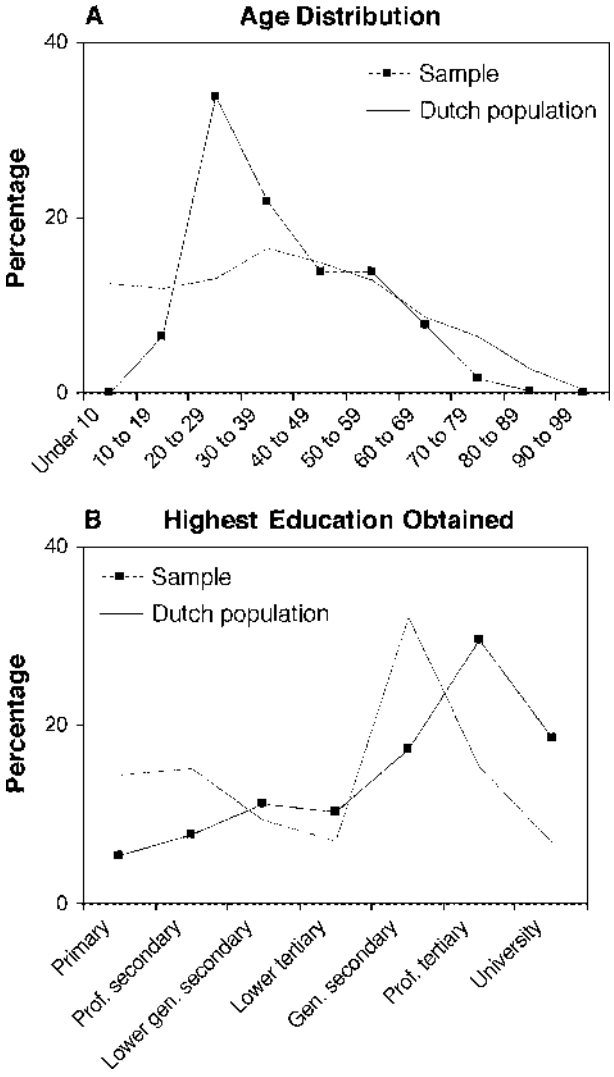

Figure 1. Distributions of (A) age and (B) education in the Dutch sample. Both are compared with the distribution of both variables in the general population of the Netherlands (source: www.cbs.nl). In accordance with statistical convention in the Netherlands, education was classified by highest attained educational grade.

Analysis

In all, 238,547 data points were included in the analyses. To ease fitting, we pooled the data into bins of 3 days. Because 60% of the presented questions were less than 2 months old, there was a large drop in the number of observations for retention intervals longer than 60 days. Retention intervals longer than 60 days were therefore pooled into 10-day bins. Bins were labeled with their middle value (e.g., the 3- to 5-day bin was labeled as having a 4-day retention interval). For 4AFC questions, the number of observations per bin was 3,208 on average, varying from 1,163 to 9,873. For open questions, the number of observations per bin varied from 585 to 4,557

Table 2
Regression of Participant Mean Scores to the Biographical Data Provided by the Participants

	Dutch		International: USA		International: Other	
Predictors	Norm. Weights	Difference Over Entire Range	Norm. Weights	Difference Over Entire Range	Norm. Weights	Difference Over Entire Range
Newspaper reading	.28	.13	.13	.05	.16	.07
Gender	.22	.06	.19	.05	.18	.05
Education	.18	.10	.20	.11	.03	.02
Televised news	.11	.05	.08	.03	.18	.08
Age	.05	.03	.11	.07	.16	.12

Note—*Norm. Weights* refers to the *b*-coefficients in the multiple-regression model after normalization of the predictors. *Difference Over Entire Range* refers to the difference between predicted performance at the minimum value of the variable from that at the maximum value. As an example, performance is predicted to be 10% higher for participants in the Dutch sample with the highest education than for those with the lowest education.

(M = 1,379). Figure 2 shows the resulting retention curves for the two formats, 4AFC and open questions.

As is customary in the literature, we have fitted curves that represent group averages. However, conclusions reached on the basis of group data do not always apply at the level of individuals, and vice versa. If functions have parameters with nonlinear effects—as is the case with forgetting curves—averaging can also yield functions that have a different form from the component functions (e.g., Brown & Heathcote, 2003; Estes, 1956). Indeed, averaging MCM retention functions with different parameter values does not necessarily yield an MCM function, and the same is true for Weibull functions. Although this is a serious caveat, fitting curves to individual data would require more data per individual than were available here.

Figure 3 shows all model variants that were fitted to the data, organized into two separate hierarchies for the model families that were considered, as well as the unrelated power model. Variants are ranked according to the number of parameters they need to fit the two functions, from eight to three. Variants with fewer than three parameters produced such bad fits that they will not be mentioned. Note that each variant is connected to more and less restricted variants: A variant will be called a *submodel* of another if it is equivalent to that other variant in all respects, except that one or more parameters have been set to a default value. The variant with fewer restrictions will be called the *supermodel* of the submodel. For the 4AFC questions, we fitted the models in Figure 3, after a correction for guessing.

Fitting was done in two steps: Because maximum likelihood estimation often veered into unproductive parts of the parameter space, we started by fitting each variant using the more robust MS_e method. The values reached by that method were taken as a starting point for maxi-

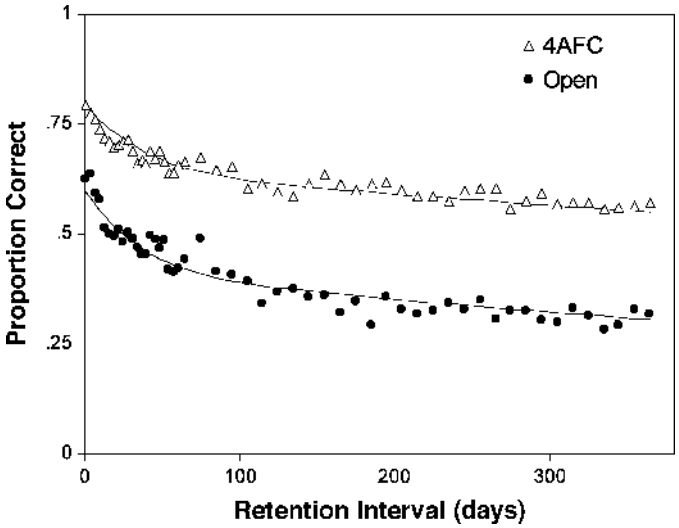

Figure 2. Retention curves for the Dutch sample (Experiment 1). Plotted separately are the data from the 4AFC and the open questions. Continuous lines represent the best-fitting MCM variant, with parameters listed in Table 3.

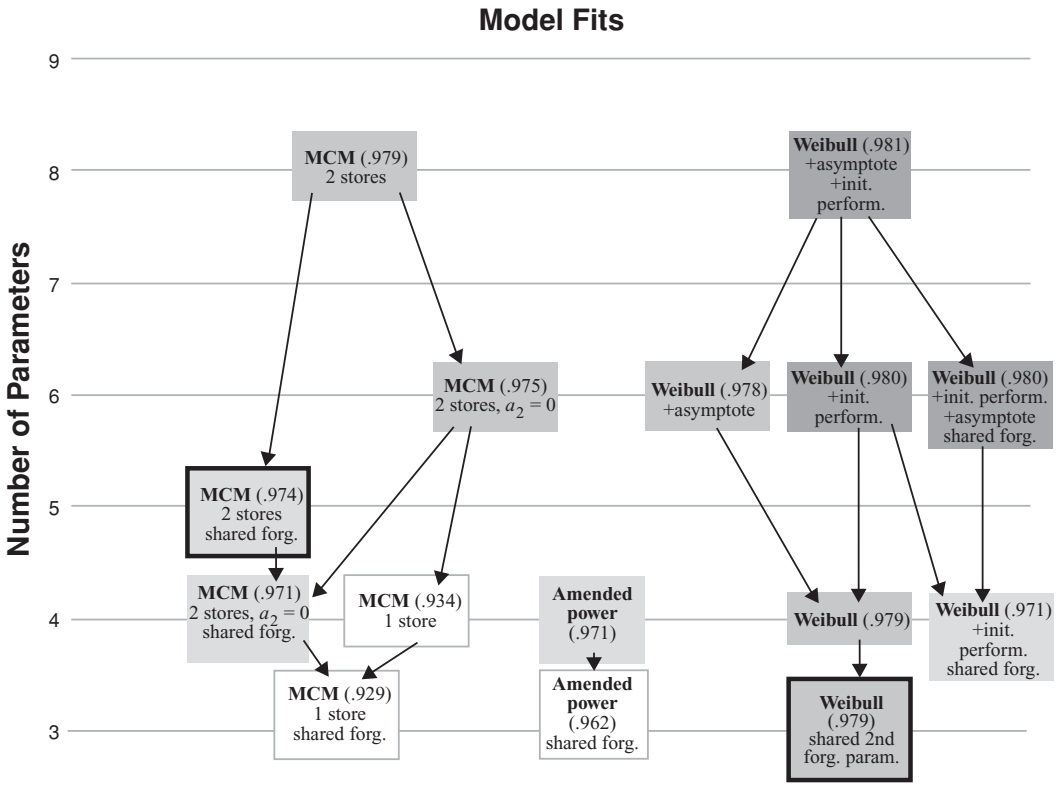

Figure 3. Comparison of all models when simultaneously fit on the open and 4AFC question data for the Dutch sample (Experiment 1). Models are ranked according to their number of free parameters, and goodness of fit is color coded (fit is also in parentheses following the model name). Supermodels are connected with their submodels by arrows. Models surrounded by black boxes were retained for further consideration; see the text for details. MCM, memory chain model; Weibull, extended Weibull model; Shared forg., decline parameters shared between recall and recognition; Init. perform., separate parameter for initial performance below 1.

mum likelihood estimation of parameter values. For each model family, we fitted variants with the most and the fewest parameters to the data by starting from several combinations of values set by hand. Following that step, we fitted variants with intermediate numbers of parameters from two starting points, the parameter values of both the supermodels and submodels.

To decide which variants were worth pursuing, we computed the BIC (Bayesian information criterion; Schwarz, 1978) for each variant. The BIC allows models to be compared while correcting for the number of parameters. It has as an advantage over similar measures (e.g., the Akaike information criterion) in that its validity does not rest on the assumption that the true model is in the same family as the ones that are being compared (Zucchini, 2000).

After examining the fits of all variants, we retained from each family the variant with the lowest BIC. For the MCM family, this was a two-store model with forgetting for the second store and shared forgetting parameters for recall and recognition (BIC = 551.7, five parameters in total). For the extended Weibull function family, it was a variant without an asymptote or an initial learning parameter and with a parameter for balance between early and late forgetting shared by recall and recognition (BIC = 504.6, three parameters). A power law model with separate parameters for both curves also did well (BIC = 545.6, four parameters). Figure 3 reports the performance of all variants (reported as R^2 for higher intuitiveness), and Table 3 lists the parameter values of the best-fitting variants.

Only the retained MCM variant predicted an immediate performance (i.e., retention at an interval of 0 days) of less than 1. This is a natural characteristic of the MCM. The retained Weibull model fitted better than a supermodel in which immediate performance was free to vary (BIC = 511.5, four parameters). The MCM predicted a performance of 79% correct for the 4AFC questions at lag 0 and 60% correct for the open questions. This result probably reflects the nature of the material: Not all news events are attended to by all participants; thus, maximum performance—lower than 1—is set by the number of participants who know of the news event in question. The retained Weibull variant, although it predicted 100% correctness for both 4AFC and open questions at lag 0, produced less than perfect performance at other short in-

Table 3
Parameter Values of the Best-Fitting Memory Chain Model and Extended Weibull Function Model for Whole Data Sets of the Three Experiments

Experiment	Question Type				
		Memory Chain Model			
		μ_1	a_1	μ_2	a_2
1	Open	0.92	.032	.018	.0010
	4AFC	1.29	.032	.018	.0010
2	Open	0.49	.018	.011	0
	4AFC	0.64	.018	.011	0
3	Open	0.76	.020	.010	.00045
	4AFC	1.20	.020	.010	.00045
		Weibull Function Model			
		μ	a	d	b
1	Open	1	.0013	.20	0
	4AFC	1	.00026	.20	.25
2	Open	1	.0089	.087	0
	4AFC	1	.0011	.087	.25
3	Open	1	.0018	.17	0
	4AFC	1	.00017	.17	.25

tervals by assuming stark forgetting in the first days after a news event.

We will concentrate our further presentation of the results on the two aspects of retention discussed in the introduction.

Final Asymptote

Both of the retained models set the asymptote for retention at 0 and predicted negligible retention ($<10^{-5}$) if extrapolated to a retention interval of 10 years. Hence, there was no evidence for an asymptote of forgetting. The MCM submodel with an asymptote (two stores, $a_2 = 0$) performed worse than the retained model, which had nonzero forgetting from Store 2 (BIC = 552.7, four parameters), and a Weibull supermodel with a nonzero asymptote parameter was also rejected (BIC = 516.9, four parameters). An amended power curve with asymptote also fitted worse (BIC = 570.4, six parameters) than one without such a parameter (BIC = 545.6, four parameters).

Shared Parameters

Recall and recognition. Correcting for guessing in the 4AFC format was not enough to fit the curves generated by the two question formats in either model family. The best-fitting MCM variant had separate learning parameters for the two formats and shared decline parameters. A variant with separate decline parameters for the open and the 4AFC curves was rejected (BIC = 566.7, eight parameters). In contrast, the best-fitting Weibull variant had a fixed initial performance parameter (set to 1) and separate decline parameters for recall and recognition. A variant in which the opposite was true—shared decline, separate initial performance parameters—had a worse fit (BIC = 556.5, four parameters).

The two model families thus account in different ways for the differences between the forgetting curves of recall and recognition. The MCM suggests that participants have an advantage when they recognize rather than generate the answer and that this advantage remains constant through time. The Weibull model points to a situation in which participants initially retrieve the answer as readily as they can recognize it, but the answer quickly becomes hard to recall while still being recognized correctly in the 4AFC test.

In the fits above, it was assumed that with the 4AFC questions, participants who did not know an answer had a 25% likelihood of guessing the right alternative. It is possible, however, that the participants could in fact eliminate one or more alternatives from consideration or that the correct alternative was more likely to be chosen than others even by participants who forgot the event in question. To investigate whether such hypotheses would explain the differences in forgetting rates between open and 4AFC questions, we compared the models described above with others in which all parameters were shared between the two formats but an additional guessing parameter was introduced for the 4AFC questions. In the Weibull framework, a variant with a free guessing parameter fitted as well as a model assuming differences in decay rates (BIC = 504.6, three parameters), and in the MCM framework such a model fitted better than the model assuming differences in learning (BIC = 533.9, five parameters). Both models estimated a probability of guessing the right answer of 31% instead of 25%. A likelihood of guessing the correct alternative higher than mere chance may thus be the most parsimonious explanation for the differences in retention of recall versus recognition. The variants that include this assumption were used in the fits reported below.[1]

Fitting differences in acquisition. Initial learning was not manipulated in our study. However, a way to investigate acquisition and its effect on retention is to look at subgroups of the population. Newspaper consumption not only is a natural proxy for the amount of learning about the news, it was also the best predictor of participant mean score. We compared the 4,236 tests finished by participants who read many newspapers (6–7 per week) to the 2,222 tests finished by participants who read newspapers sporadically (0–2 per week). Retention intervals were again pooled into 3-day bins for intervals under 60 days and 10-day bins for intervals above 60 days. The four curves defined by newspaper reading level and question format were fitted with the two retained models (see Table 4). Figure 4 shows the resulting curves.

The MCM had a better fit when the curves for the two participant groups were fitted with shared decline parameters than when separate decline parameters were used (shared decline, BIC = 998.4, 6 parameters; separate decline, BIC = 1,011.9, 10 parameters). The same was true for the Weibull model (shared decline, BIC = 937.0, 5 parameters; separate decline, BIC = 972.5, 8 parameters). Retention could thus be separated from initial learning level, which was higher for regular than for

Table 4
R^2s of the Best-Fitting Memory Chain Model (MCM) and Extended Weibull Function Model for the Different Data Sets

			Readers vs. Nonreaders				Older vs. Younger Adults			
	All Data		Nonshared Forgetting		Shared Forgetting		Nonshared Forgetting		Shared Forgetting	
Model	R^2	FP	R^2	FP	R^2	FP	R^2	FP	R^2	FP
					Experiment 1					
MCM	.974	5	.971	10	.969	7	.942	10	.940	7
Weibull	.978	3	.973	6	.973	4	.934	6	.923	4
					Experiment 2					
MCM	.964	4					.831	8	.817	6
Weibull	.960	3					.821	6	.815	4

Note—FP, number of free parameters used to fit the data.

sporadic newspaper readers in both the MCM and Weibull models. The data could be accounted for even better in the MCM if the two groups were given separate guessing parameters in recognition (BIC = 992.9, 7 parameters), suggesting that more knowledge allowed frequent newspaper readers to eliminate slightly more options in the 4AFC format (guessing 41% correct, in contrast with 36% for the infrequent newspaper readers).

Participant age. Another participant variable that can be investigated is age. Effects of age on retention, with older adults exhibiting faster forgetting, have been found by some researchers (e.g., Brainerd et al., 1990; Wheeler, 2000) but not by others (Rubin & Wenzel, 1996). We therefore compared all participants older than 60 years of age with participants between 18 and 24, which led to sample sizes of 1,371 older and 1,168 younger adults. Figure 5 shows the retention curves of both groups for the open and the 4AFC items.

Again, 3-day bins were used for retention intervals of up to 60 days and 10-day bins for longer retention intervals. The four curves defined by the two question formats and two age groups were fitted with the retained models as above. Fits were worse when decline parameters were shared by the two groups than when they were not for both the MCM (shared decline, BIC = 879.3, 6 parameters; separate decline, BIC = 842.6, 10 parameters) and the Weibull function (shared decline, BIC = 862.6, 5 parameters; separate decline, BIC = 819.9, 8 parameters). Both models pointed to somewhat steeper forgetting for older than for younger adults. However, fits were best when forgetting was assumed to be equal for both groups, but the likelihood of guessing the right answer in the 4AFC condition varied between the two groups (BIC for MCM = 817.8; BIC for Weibull = 802.5). Both models set the likelihood that older adults guessed the right answer higher than the likelihood that younger adults guessed the right answer, in line with the older adults' higher learning parameter in both models.

Power analysis and controls. In order to investigate whether the findings reported above were influenced by a lack of power, we studied how much of a change to either the learning or the main decline parameter from the optimal model would compensate for the loss of one parameter in either the MCM or the Weibull model. It turned out that a 3% decrease or increase in the value of the learning parameter led to a reduction in the BIC equivalent to one parameter, as did a 4% increase or 3% decrease in the main decline parameter. In the Weibull model, a 2% decrease or increase in the learning parameter or a 4.5% decrease or increase in the most critical decline parameter led to the model being rejected against the model with the original parameters. Small changes in parameter values thus already lead to rejection of the models, indicating that the results obtained above were not due to a lack of power.

Theoretically, participants could have answered questions by looking up the answers on the Internet. Although there would be no reward for such cheating, it cannot be excluded that some participants engaged in it. To ascertain that such activity would not have a large impact on our results, we performed two checks: First, we searched for individuals with a perfect score, who might have been suspect. None of our participants had one, however. Second, we reasoned that an Internet search must have taken some time, and that therefore cheating might reveal itself in answers with low latencies. We therefore computed an estimate of reaction time by comparing time stamps of responses in our database. Although such times also include transportation times to and from a participant's computer, they form a rough estimate of how many seconds a participant spent on a particular item. There was no strong correlation, however, between this reaction time and likelihood of a correct answer.

In order to test this proposition formally, we divided our data set into trials on which participants had spent more than 12 sec (plus an estimated 3 sec for Internet delays) and trials on which they spent 12 sec or less. There was no difference in proportions of correct answers between fast and slow answers on the open questions [$t(6,719) = 0.71, p < .48$]. There was a slight difference on the 4AFC questions [$t(7,628) = 2.01, p = .045$], but faster responses had the advantage, with a mean 56.7% versus 56.1% correct, respectively, contradicting the hypothesis that slow responses benefited significantly from

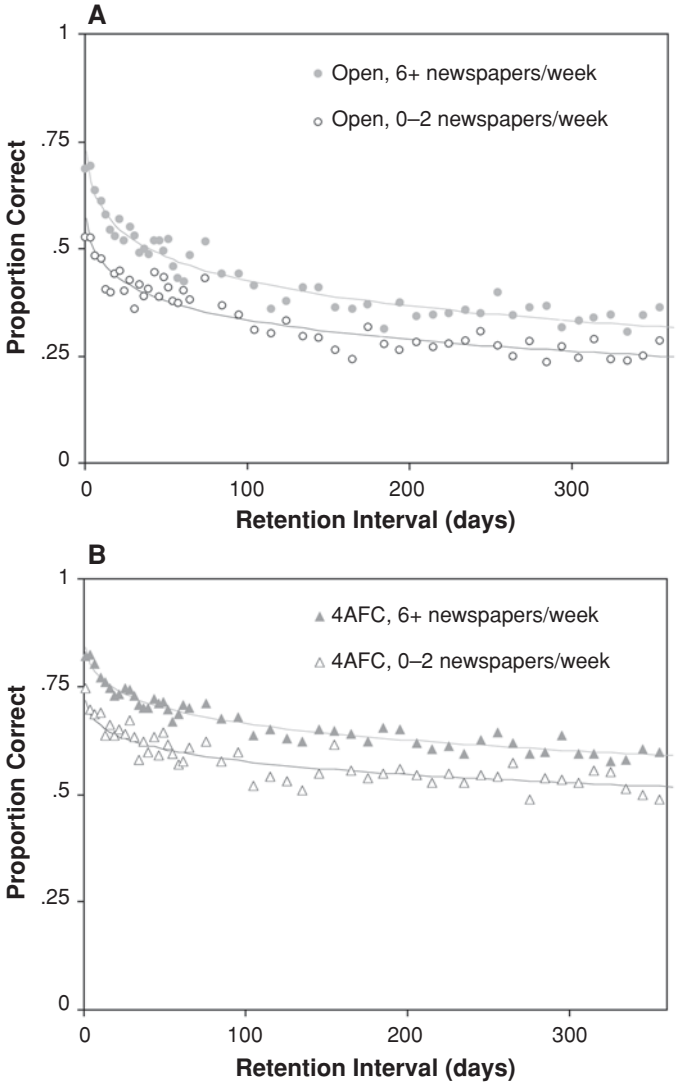

Figure 4. Open (A) and 4AFC (B) retention curves for the Dutch sample (Experiment 1). Plotted separately are those participants who read many newspapers (at least 6 a week) and those who read few newspapers (0–2 per week).

cheating. For both formats, the retention curves of the fast and slow answers were virtually on top of each other.

Discussion

Fits with the two model families were mostly in agreement. Both the Weibull and MCM frameworks suggested that recall and recognition have the same decline function, with the caveat that general knowledge aids recognition, so that the likelihood of guessing the right answer was higher than mere chance. In support of this conclusion, it was found that the groups that started out at a higher level of performance (consumers of the news and older adults) had a higher guessing parameter than did those who started out at a lower level of performance (participants who read fewer newspapers and college-age adults). In both of these comparisons, equivalent decline parameters were found for the two performance groups, supporting the conclusion that forgetting can be dissociated from the initial level of learning/performance.

Moreover, both models suggested that there was no asymptote in retention, or at least not in the recall question format. The guessing likelihood parameter introduced in both models may have functioned as an asymptote for the 4AFC format. Since the form of retention was the same for the recall and 4AFC formats, however, it is unlikely that this parameter can be interpreted as an asymptote only to be found for recognition.

In one instance, the two model frameworks did not lead to the same conclusion—namely, on the question of the initial level of performance. The MCM has as a characteristic that retrieval is probabilistic, leading to suboptimal performance at lag 0 even when an item has been

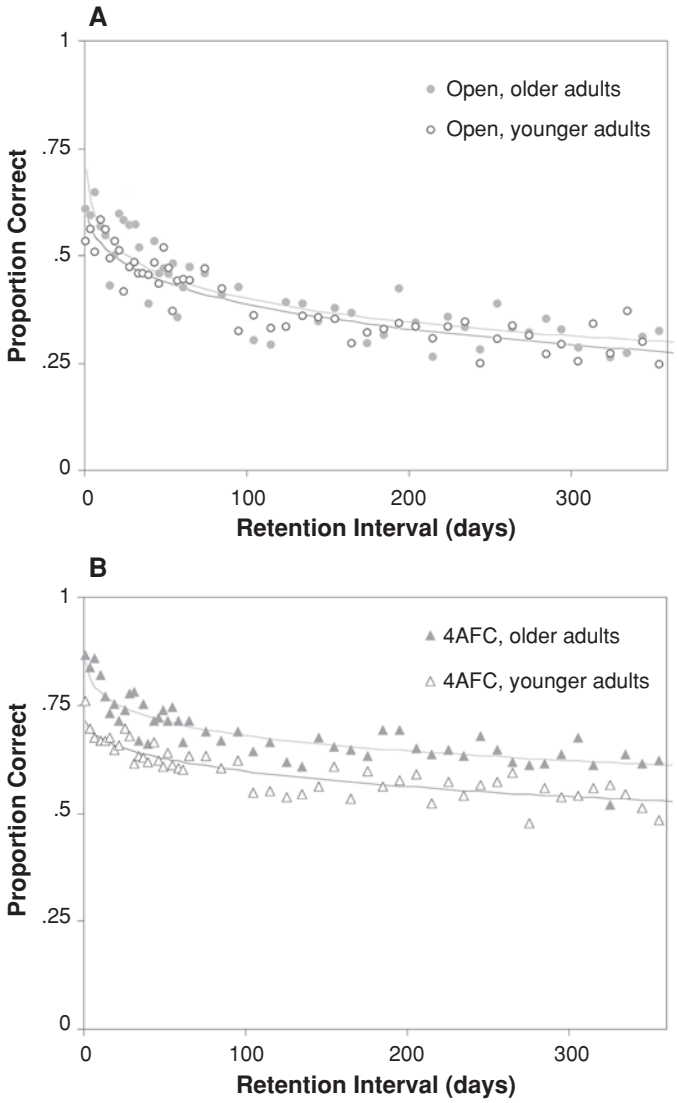

Figure 5. Open (A) and 4AFC (B) retention curves for the Dutch sample (Experiment 1). Plotted separately are older (age 60 or greater) and younger (ages 18–26) participants.

well learned. Although the Weibull framework allows performance to start off at a level below 1, variants including this feature did not improve the fit of the model. Because our news test covered many categories of news, it is highly unlikely that each participant had originally been exposed to every item, making a theoretical initial performance of 100% extremely unlikely. In that respect, the Weibull model's behavior here is not entirely satisfactory.

EXPERIMENT 2

Method

Design of the Test

The second experiment concerned questions in English for an international audience. On February 15, 2001, an English-language version of the DNMT was opened to the public. It had the same form as the Dutch version, and questions were mostly translations from Dutch questions pertaining to international news, although some questions were taken from headlines at Internet sites dedicated to international news. In all, 418 usable English-language questions were formulated. Another 50 were not used because participants scored below chance performance level on them in a multiple-choice format. Participants were asked to give the same information about themselves as in the Dutch test. Since there is no widely known international system to code educational achievement, participants were asked how many years of formal education they had completed and also to list their country of residence.

Participants

The international version of the test was completed 9,657 times by 7,149 participants (only 19% of the international participants performed the test more than once). About 50% of the participants originated from the United States. Other primarily English-speaking

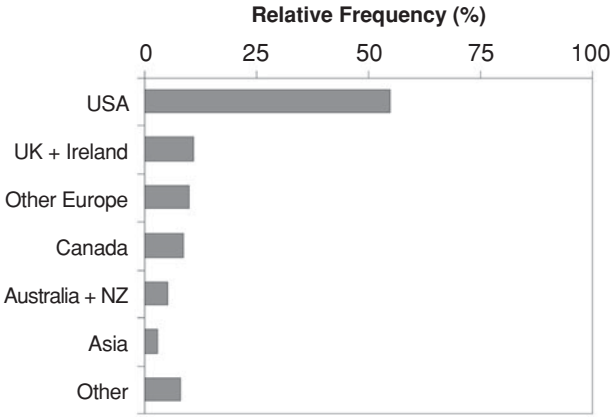

Figure 6. Country of residence of participants in the international sample (Experiment 2).

countries were also well represented (e.g., United Kingdom, Canada, and Australia; see Figure 6). As with our Dutch sample, the international participants were comparatively young and well educated (see Figure 7). Of the international participants, 46% were male and 54% female.

Results

Performance and Predictors

Performance on the international test was worse than on the Dutch test, with participants answering correctly only 31% of the open questions and 52% of the 4AFC questions on average. Regression analyses on participant means were done separately for those who identified the United States as their country of residence and those who identified other countries (see Table 2). For American participants, education and age were the best predictors of mean score, but for the remaining participants all variables except education were approximately equally powerful predictors (educational systems probably vary too much from country to country for years of formal schooling to be a good predictor). Because of the differences between the American participants and those from other nations, we decided to use only data from the American participants in our analyses of forgetting; this decision reduced a potentially large source of variability and still left us with the majority of our data (5,086 of the finished tests).

Analysis

In all, 158,476 data points were included in the analyses. As with the Dutch sample, we grouped the data into 3-day bins for retention intervals of up to 60 days and 10-day bins for longer retention intervals. Figure 8 shows the resulting retention curves.

The two curves (for open questions and 4AFC questions) were fitted with the same variants of the MCM and extended Weibull model displayed in Figure 3. Again, we computed the BIC of each variant in order to determine which ones described the data best. For the Weibull family, this model was the same one retained in Experiment 1, a model with neither initial learning nor asymptote parameters and with the parameter determining the balance between early and late forgetting shared between recall and recognition (BIC = 491, three parameters). The best MCM variant was different, however: a two-store model that, unlike the model retained in Experiment 1, had no forgetting from Store 2 (BIC = 493.7, four parameters). An amended power curve was also fit to the data (BIC = 495.2, four parameters).

In large part, conclusions from the fitting were similar to those reached from analyzing retention of the Dutch questions. Again, the retained MCM variant predicted an immediate performance (i.e., retention at lag 0) of less than 1, but the retained Weibull model had immediate performance set at 1. The MCM predicted a performance of 60% correct for the 4AFC questions at lag 0 and 39% correct for the open questions.

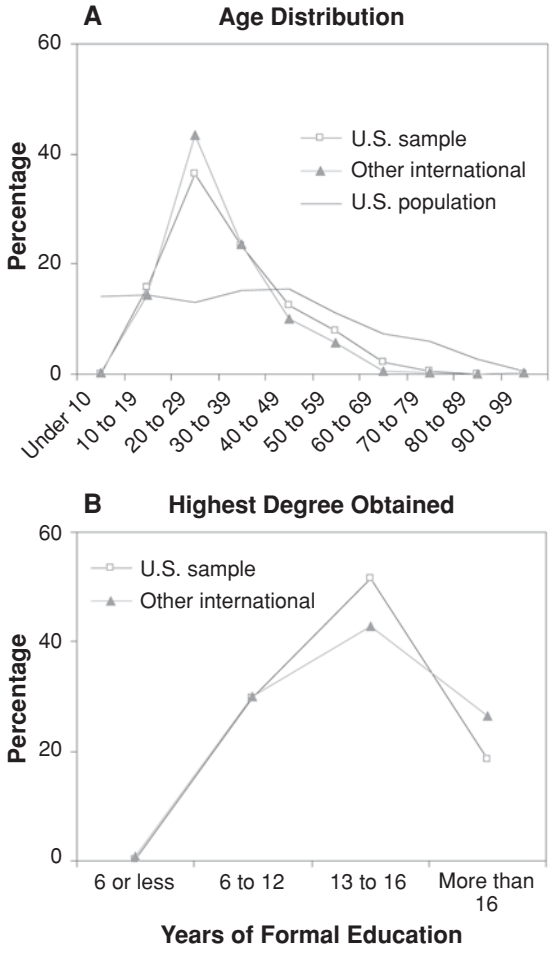

Figure 7. Distributions of (A) age and (B) education in the international sample (Experiment 2). Participants originating from the United States are plotted separately from those originating from other countries. For age, an estimate of U.S. population distribution was added (source: www.census.gov). Education was elicited as the number of finished years of formal education.

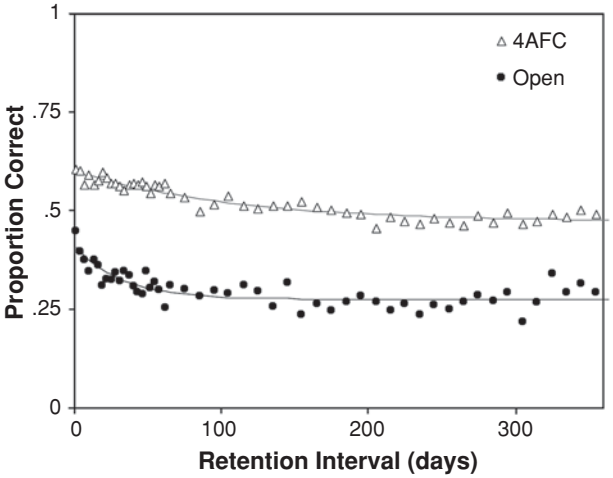

Figure 8. Retention curves for the international sample (Experiment 2). Plotted separately are the data from the 4AFC and the open questions.

Final Asymptote

A Weibull model with a nonzero asymptote parameter was rejected (BIC = 501.2, four parameters), as was also the case with the Dutch sample. The same was true for an amended power curve with an asymptote (BIC = 518.4, six parameters). The retained MCM variant, however, with consolidation to a second store from which no forgetting occurred, performed better than a supermodel with nonzero forgetting from Store 2 (BIC = 505.7, five parameters). The best-fitting MCM variant thus predicted an asymptote. Moreover, the best-fitting Weibull variant predicted such slow forgetting that performance after very long intervals was not much worse than what the MCM predicted: After 10 years, performance on the open questions was predicted to be 26.3% by the MCM, as compared with 17.7% predicted by the retained Weibull variant (and 21% by the amended power curve without asymptote).

Shared Parameters

Recall and recognition. For the MCM framework, the retained variant with shared decline parameters for recall and recognition but separate initial learning parameters fitted better than a variant with separate decline parameters for the two formats (BIC = 503.2, six parameters). The retained Weibull variant had separate decline parameters for recall and recognition but a shared initial learning parameter. It fitted better than a variant in which the opposite was true (BIC = 518.1, four parameters). These were similar to the results in Experiment 1. Unlike in Experiment 1, however, assuming shared decline parameters but a likelihood greater than .25 for guessing the correct answer to 4AFC questions did not provide an equally good description of the data; such a variant had a worse fit in both the MCM (BIC = 507.2, four parameters) and Weibull (BIC = 505.3, four parameters) frameworks.

Participant age. Since newspaper reading was not a strong predictor of score in the international sample, no attempt was made to divide the sample into two groups based on this variable. Age, however, was analyzed similarly to how it was in Experiment 1. Using the same definitions from Experiment 1, we obtained samples of 251 older adults (ages 60 and older) and 1,346 younger adults (ages 18–24).

Figure 9 shows the resulting retention curves. Again, we fitted these curves with the two retained models and tested whether the data could be fitted, assuming shared decline parameters for the two groups. Indeed, fits were better when decline parameters were assumed to be shared between the older and younger groups. This was the case both for the MCM framework (shared decline, BIC = 770.6, 6 parameters; separate decline, BIC = 773.9, 10 parameters) and the Weibull function (shared decline, BIC = 747.5, 4 parameters; separate decline, BIC = 755.0, 6 parameters). Both models set initial perfor-

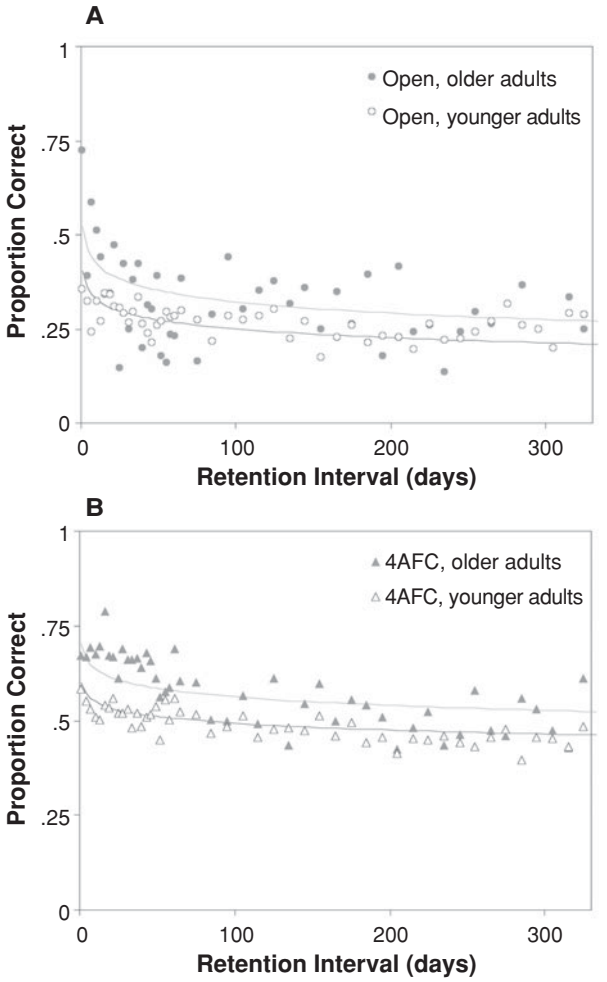

Figure 9. Open (A) and 4AFC (B) retention curves for the American participants in the international sample (Experiment 2). Plotted separately are older (age 60 or greater) and younger (ages 18–26) participants.

mance a little higher for the older adults than for the younger adults.

Discussion

In some respects, the results of Experiments 1 and 2 were in harmony; notably, the same variants of both frameworks provided the best overall fit of the data, and in both data sets older and younger adults showed the same decline function, even though initial performance of the two groups was dissimilar. In two ways, however, the results of the American participants in our international sample are different from those in our Dutch sample. Although no asymptote in performance was found in the Dutch sample, the MCM fits in Experiment 2 did point to one (the Weibull fits also pointed to such slow decline that it was difficult to distinguish from an asymptote in forgetting). Also, in Experiment 1, a likelihood of guessing the correct answer higher than chance level explained the difference between the retention curves for recall and recognition, but in Experiment 2, this was not the case. This difference is convolved with the issue of whether there is an asymptote in forgetting, since the guessing parameter functions as the asymptote in the retention curve associated with the 4AFC format.

EXPERIMENT 3

To further investigate the issue of an asymptote in forgetting, we ran a third experiment in which retention was tested over a 2-year rather than a 1-year period.

Method

Design of the Test

In Experiments 1 and 2, questions were sampled with a 30% likelihood from the 30 days before the test, a 30% likelihood from approximately a month prior to that (31–60 days before the test day), and a 40% likelihood from before that. From February 28, 2003, the composition of the test was changed by sampling questions from five periods instead of from three. Twenty-five percent of the questions were now sampled from the most recent month, another 25% from the next-to-last month, and the remaining questions from three periods: 30% from the period between 61 and 365 days before the test (i.e., questions that were up to 1 year old), 10% from between 518 and 548 days before the test (i.e., questions that were around 1.5 years old), and 10% from between 700 and 730 days before the test (i.e., questions that were around 2 years old). The rest of the procedure was the same as in the previous experiments.

Participants

The new format was introduced for both the Dutch and international versions. However, because the number of international participants had by then dropped dramatically, we restricted ourselves to analyzing retention for participants using the Dutch version (only 313 tests were finished by participants from the U.S., which was too few for reliable analyses). From February 28, 2003, to February 14, 2004, the Dutch version of the test was completed 3,956 times by 2,853 participants.

Results

Performance and Analysis

Performance in the third experiment was comparable to that in Experiment 1, although it was a little lower because of the longer retention intervals. Participants correctly answered on average 37.5% of the open questions and 64% of the 4AFC questions.

In all, 116,095 data points were included in the analyses. We again pooled data into 3- and 10-day bins; each bin contained from 177 to 2,682 data points. The resulting retention curves are shown in Figure 10.

Fitting the curves with the MCM and the Weibull function led to the same variants being retained as in Experiment 1. For the MCM family, this variant was a two-store model with forgetting from the second store and

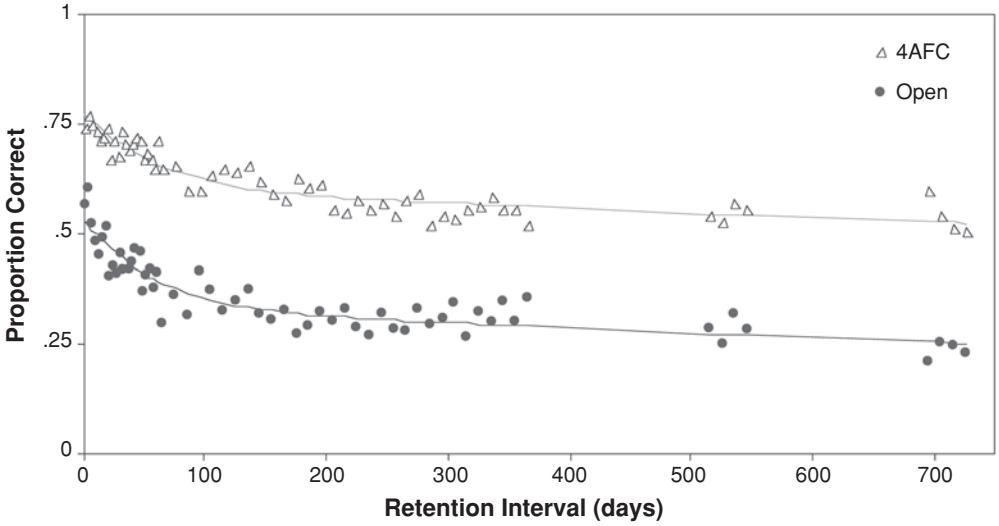

Figure 10. Two-year retention curves for the 4AFC and the open questions for the Dutch sample (Experiment 3). Continuous lines correspond to the fits of a two-store MCM with decline parameters shared by 4AFC and open questions, with parameter values fitted only on the first year of retention.

shared forgetting parameters for recall and recognition (BIC = 621.8, five parameters). For the extended Weibull function family, it was a variant without asymptote or an initial learning parameter and with the parameter for the balance between early and late forgetting shared by recall and recognition (BIC = 627.0, three parameters).

Final Asymptote

Neither of the retained models showed an asymptote in forgetting. Introducing an asymptote worsened the fit for both the Weibull model, in which it entailed introduction of an extra parameter (BIC = 636.7, four parameters), and the MCM, in which it entailed an extra restriction (BIC = 628.1, four parameters).

In order to investigate the possibility that the asymptote is only evident for retention intervals longer than 1 year, we fitted the data of the first year in isolation. We then investigated whether parameters found in that way would underestimate retention in the second year. If this were the case, it would be evidence for an asymptote appearing late in retention. In fact, adjusting parameters to fit only the first year of retention did not noticeably detract from the fit for the entire 2 years for either the Weibull framework (BIC for the fit on the entire 2 years: 627.1, three parameters) or the MCM (BIC for the fit on the entire 2 years: 623.7, five parameters). The lines shown in Figure 10 are those of the MCM variant with parameter values adjusted to fit only the first year of retention. As can be seen, it also fits the second year of retention well, leading neither to systematic under- nor overestimation of retention. This result supports the conclusion from Experiment 1 that there is no asymptote in retention for the Dutch version of the test.

GENERAL DISCUSSION

In this article, we presented a study in which retention of news events was tested for intervals ranging from 1 day to 2 years. Two versions of the test, one in Dutch and one in English, were analyzed separately but led to essentially the same conclusions.

The best-fitting MCM variants pointed to a performance at a retention interval of 0 ranging from 74% for the Dutch 4AFC questions (Experiment 1) to 46% for the international open questions (Experiment 2). This range may correspond to the proportion of the Internet population that is exposed to major headline items in the news. After initial exposure, proportions correct dropped to around one third of these values at the longest retention intervals (335 days) on open questions and to around two thirds at similar intervals on 4AFC questions.

A feature that distinguishes the present retention study from previous ones is its use of the Internet as a means to deliver a memory test to participants. Although not many people would dispute that the Internet is a useful tool in scientific research, few studies so far have used it to generate actual data. Several possible confounds that Internet data gathering may introduce have been discussed in the literature (Buchanan & Smith, 1999), but what comparative research exists has tended to show a general equivalence between data gathered via the Internet and via traditional methods. For example, Buchanan and Smith found that data from a personality test taken by a sample of Internet volunteers had the same psychometric characteristics as data from a similar paper-and-pencil test taken by a standard sample of psychology undergraduates. Even reaction time experiments delivered over the Web have elicited the same experimental effects as similar experiments in a laboratory setting (McGraw, Tew, & Williams, 2000).

The most serious drawback of Internet research is probably the lack of representativeness of Internet samples, both because of the characteristics of the Internet population and because Internet participants volunteer their time (Buchanan & Smith, 1999). Internet users tend to be younger, better educated, and predominantly male (except for the last point also a good description of psychology undergraduates). Moreover, their volunteering may imply a high motivation and interest in the topic of the research. In our study, however, since retention interval was manipulated within participants and performance in absolute terms was not important, possible sampling errors were not relevant. In addition, the possible disadvantages of research over the Internet were more than outweighed in our case by the possibility of including a very large number of participants. In all, more than 14,000 participants took part in this study, allowing for a wide range of questions surrounding retention and forgetting to be addressed and giving the analyses great power.

The fact that participants did not study the material to remember in a laboratory setting squarely places this research in the tradition of research with naturalistic retention material (Rubin & Wenzel, 1996). In this tradition, the materials have the disadvantage of the study schedule typically being unknown. In the case of news events, participants may be confronted with a particular news event for weeks in television news or newspapers. However, an attempt (not reported) to model relearning explicitly did not yield any increase in fit. An analysis of the difference between retention curves produced by models with and without relearning pointed to media coverage concentrated in the days immediately following the news event.

The methodology followed here in analyzing data relied on the fitting of models to the data. The two models with which the fitting was done both had four free parameters per curve, making them rather flexible. Flexibility in models has been justly criticized, since flexible models may fit anything and nothing without leading to a deeper understanding of the mechanisms behind the fitted curves (Roberts & Pashler, 2000). However, the models that were found to fit best had a small number of parameters, because some parameters proved to be unnecessary to describe the data and others were shared by different curves. Moreover, models were fit to the data

not to support the models themselves, but as a means to test hypotheses concerning retention. Our conclusions with reference to the hypotheses outlined in the introduction will be reported below. It bears repeating, however, that these conclusions were reached on the basis of group data only. It is possible that a different picture would emerge if data from individual participants were fitted separately and analyses were done over parameter values, but such an analysis would have required more data per participant than was available here.

Asymptote

One aspect of retention concerns the final asymptote of performance. For the Dutch test, both models predicted a zero asymptote (Experiments 1 and 3). For the international test, the MCM predicted a nonzero asymptote, whereas the extended Weibull function predicted instead very slow forgetting (Experiment 2). The nonexistence of an asymptote in memory for news would be inconvenient for constructors of retrograde amnesia tests. These tests rely on news events as material and would have to be renormed every few years if forgetting did indeed take place after long intervals. Moreover, since an asymptote has been found in other memory domains (Bahrick, 1984, 1992; Bahrick et al., 1975; Bahrick & Phelps, 1987; Rubin et al., 1999), its nonoccurrence in Experiments 1 and 3 is more puzzling than its occurrence in Experiment 2.

One explanation for the inconsistent results in this study and between our Dutch data set and other studies is that an asymptote in forgetting may only exist for overlearned material and not for the relatively detailed facts on which most of the questions in the DNMT were based (see the Creation of the Questions section in Experiment 1). With the exception of Rubin et al.'s (1999) study, studies finding an asymptote have tended to rely on material, such as school English or faces, that has been rehearsed many times. Here, international questions were selected from the Dutch corpus for translation for their relevance on a world level, and therefore may have had topics that were most likely to be news events of lasting importance. These questions may thus have been the most likely to be overlearned. However, this explanation is only speculative, and other aspects of our study may also explain our results. In particular, it is possible that a 2-year interval is not long enough to spot the emergence on an asymptote in retention.

Shared Parameters

The influence of three variables on parameter values was investigated in this article.

Recall and recognition. One may assume that recognition and recall are based on the same memory store and that better cuing in the recognition format is the only difference (because memory is cued with the item itself). In the MCM, cuing effects are incorporated in a parameter that also reflects the strength of initial learning, and fits of the MCM were indeed best in all data sets when only this parameter varied between recall and recognition. In the Weibull framework, no clear way exists to incorporate such cuing effects. Instead, the difference between the two question formats was covered by different decay parameter values, with stronger decay for open questions. These parameter values imply that participants at the outset can answer a question about a certain event as well in open form as in 4AFC form and that their ability to reproduce the answer for an open question subsequently declines faster than their ability to recognize the correct answer in a 4AFC question.

Both of these pictures—of a cuing benefit for recognition throughout retention or of faster decay of the ability to reproduce versus recognize—are appealing. On the surface, forgetting seems steeper for open questions than for 4AFC questions, especially at short retention latencies, replicating previous studies (Rubin & Wenzel, 1996). In the MCM, however, a stronger early drop in performance for open questions is a natural consequence of a lower learning/cuing parameter and not the reflection of genuinely steeper decline. More studies on this topic are probably called for.

Less steep forgetting for the 4AFC questions may also be an artifact created by guessing. If the likelihood of guessing the correct answer—even without knowledge of the news event in question—is larger than the chance level of 25%, then this is equivalent to a higher asymptote in performance, which in turn would make the forgetting curve shallower. This explanation received support in Experiment 1 but not in Experiment 2.

Degree of learning. An old issue in the study of retention is whether forgetting is independent of the level of initial learning. Here, this question was investigated by subdividing participants into those with high and low media consumption. On average, participants who read many newspapers did not exhibit faster or slower forgetting than did participants who read few newspapers. All differences between the two data sets were found to reside in parameters identified with initial learning. This implies that forgetting is independent of the degree of learning, replicating the findings of several other studies (Rubin & Wenzel, 1996).

Age. Another variable that was investigated was age, which we operationalized by comparing college-age participants with those older than 60 years of age. In both data sets, no difference in forgetting was found between the averaged retention curves of the two groups. These results were a little ambiguous for the 4AFC questions in Experiment 1, however, because in the best-fitting model older adults were given a higher likelihood than younger adults of guessing the correct alternative. This difference led to somewhat less steep forgetting, but may have only reflected their better general knowledge.

This study is not the first not to find effects of age on retention (Rubin & Wenzel, 1996), but the consensus still seems to be that older adults forget faster than younger adults (Wheeler, 2000). One factor that may explain our results is the length of retention intervals, which were

longer than those used in most studies of forgetting. However, studies in which retention intervals have varied over a relatively wide range have tended to find stronger effects of aging over the longer intervals (Parks, Royal, Dudley, & Morell, 1988). Other factors that could play a role are educational attainment and the material used.

In many studies, age is confounded with educational attainment, because university undergraduates are compared with older adults sampled from the population (Brainerd et al., 1990; Wheeler, 2000). Here, younger adults also had a higher level of educational attainment than older adults did, but the younger adults were less educated than has been typical in studies (not all were at the university level), and the older adults were better educated than a standard sample might be. Perhaps, as a consequence, older adults started out at a slightly higher level of performance, whereas in the typical study older adults start off at a lower level of acquisition. The latter circumstance may lead to interpretational problems (Loftus, 1985) that were thus avoided in the present study.

Perhaps the most likely explanation of normal forgetting among our older adults concerns the material used. In many studies, the material has consisted of lists of words or pictures acquired in one experimental session, but here it consisted of more semantic facts acquired under naturalistic conditions. One study employing similarly semantic material also did not find evidence of faster forgetting (Cohen et al., 1992). Only for material that was highly specific did they find a small decrement in older adults, suggesting that perhaps older adults only forget material faster that has not been encoded very deeply (for further evidence, see Einstein, McDaniel, Manzi, Cochran, & Baker, 2000).

Conclusion

Rubin and Wenzel (1996) stated that comparisons of parameter values in different conditions depend on the models in which the parameters feature. The example given was that of forgetting rates of older adults that were equal to those of younger adults if retention was fitted with the power curve, but lower than those of younger adults if the logarithmic function was used to describe retention. To escape such dependency of conclusions on the model used, we analyzed our data using two, and sometimes three, models. By and large, however, the results of these tests using different models led us to the same conclusions.

REFERENCES

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.

Bahrick, H. P. (1992). Stabilized memory of unrehearsed knowledge. *Journal of Experimental Psychology: General*, **121**, 112-113.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, **104**, 54-75.

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over eight years. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 344-349.

Bogartz, R. S. (1990). Learning–forgetting rate independence defined by forgetting function parameters or forgetting function form: Reply to Loftus and Bamber and to Wixted. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 936-945.

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**, 1-92.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, **35**, 11-21.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, **90**, 125-144.

Chessa, A. G., & Murre, J. M. J. (2002). *A model of learning and forgetting: I. The forgetting curve* (Tech. Rep. No. 02-01). Amsterdam: University of Amsterdam, Neural and Cognitive Modeling Group.

Cohen, G., Stanhope, N., & Conway, M. A. (1992). Age differences in the retention of knowledge by young and elderly students. *British Journal of Developmental Psychology*, **10**, 153-164.

Ebbinghaus, H. (1885). *Über das Gedächtnis* [On memory]. Leipzig: Dunker.

Einstein, G. O., McDaniel, M. A., Manzi, M., Cochran, B., & Baker, M. (2000). Prospective memory and aging: Forgetting intentions over short delays. *Psychology & Aging*, **15**, 671-683.

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

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

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of Web-delivered experiments: Can you trust the data? *Psychological Science*, **11**, 502-506.

Meeter, M., Murre, J. M. J., & Janssen, S. M. J. (2005). *Measuring a short-term Ribot gradient: The Daily News Memory Test*. Manuscript submitted for publication.

Parks, D. C., Royal, D., Dudley, W., & Morell, R. (1988). Forgetting of pictures over a long retention interval in young and older adults. *Psychology & Aging*, **3**, 94-95.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, **107**, 358-367.

Rubin, D. C., Hinton, S., & Wenzel, A. E. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1161-1176.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

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, & Cognition*, **9**, 384-397.

Wheeler, M. A. (2000). A comparison of forgetting rates in older and younger adults. *Aging, Neuropsychology, & Cognition*, **7**, 179-193.

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

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996). *Psychological Review*, **105**, 379-386.

Wickens, T. D. (1999). Measuring the time course of retention. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson–Shiffrin model* (pp. 245-266). Mahwah, NJ: Erlbaum.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41-61.

NOTE

1. The Fisher information matrix suggests that asymptote parameters and the guessing likelihood have a negative covariance. This means that a higher guessing likelihood can partially obscure an asymptote in forgetting. However, the conclusion reported above that there was no asymptote was reached without a free guessing parameter. (Fits in the next experiment will also not put asymptote parameters against guessing parameters.)

APPENDIX

An obvious characteristic of the forgetting curve is that it is usually monotonously decreasing in time with a decreasing slope (i.e., the retention function must have a negative derivative but a positive second derivative). Wickens (1999) outlined as a further consideration that the hazard function of a plausible retention function must decrease with time. The hazard function describes the likelihood that a memory that has survived until time t will be forgotten at that time (mathematically, the first derivative divided by the raw function). A decrease in this likelihood with time implies that the older a memory is, the less likely it is to be forgotten.

Several simple functions with these characteristics have been proposed as candidate retention functions. One of these was used in this study to fit our retention curves. The power function has a long history in forgetting research and is among the most successful two-parameter functions, fitted by Rubin and Wenzel (1996) on hundreds of data sets. In its traditional form, it is ill behaved at very short retention intervals when used to fit the likelihood of a correct answer (p_{correct}): The function goes to infinity at values close to zero, whereas p_{correct} can only vary between 0 and 1. With a slight change, it starts at 1 and is still a good descriptor of forgetting (see Table 1). Other functions that are well behaved over the whole scale of retention intervals are the retention function resulting from the memory chain model and the Weibull function championed by Wickens (1999).

Memory Chain Model

One difficulty in fitting p_{correct} is the relationship between this measure and the underlying memory strength. This relationship has been explicitly modeled in the memory chain model (Chessa & Murre, 2002).

Chessa and Murre (2002) proposed that underlying memory strength may be modeled as a number of points in memory, recovery of which would lead to the correct output. These points may be copies of the memory or may be stored details that, if remembered, could trigger retrieval of the correct answer. Once created, these points are subjected to a Poisson death process, leading to a very simple model of memory strength (or, in the vocabulary of Poisson point processes, *intensity*): the expected initial number of points μ multiplied by an exponential retention function governed by decline parameter a:

$$r(t) = \mu e^{-at}. \tag{1}$$

The formula above represents the one-store memory chain model. The model assumes that memory consists of a number of stores whose dynamic is described by the equation above. Memories first reside in one store, from which they are copied to the next, and so forth. Memories may be copied from sensory registers to short-term memory, from there to a hippocampal long-term memory, and from there into a neocortical memory. In most fits, Chessa and Murre (2002) used a two-store model, and we will restrict this discussion to that version. In the two-store model, acquisition of the memory places μ_1 points in Store 1, from which they decay with a constant likelihood a_1. As long as the points exist, they may be copied to Store 2 with a constant likelihood of μ_2. From this store, they are lost with a likelihood a_2. This model results in the following function for the intensity (expected number of memories) at any time point t.

$$r(t) = \mu_1\left[e^{-a_1 t} + \frac{\mu_2}{a_2 - a_1}\left(e^{-a_2 t} - e^{-a_1 t}\right)\right]. \tag{2}$$

The derivation of this formula can be found in Chessa and Murre (2002). Retrieval is an inherently stochastic process; even if several points are available in memory, they may not be recovered. The likelihood of successful retrieval equals one minus the chance of no retrieval, which is equal, according to the Poisson distribution, to the natural exponent of minus the intensity at time point t (expected number of memories). The resulting retention function is

$$p(t) = 1 - \exp\left\{-\mu_1\left[e^{-a_1 t} + \frac{\mu_2}{a_2 - a_1}\left(e^{-a_2 t} - e^{-a_1 t}\right)\right]\right\}. \tag{3}$$

The likelihood of recovery of a point is a parameter of the model (for which the letter q is used). As q and μ_1 scale against each other, they can be subsumed into one parameter (which is also called μ_1). The μ_1 parameter thus accounts for factors in both learning and retrieval.

Weibull Function

The basic Weibull forgetting function is

$$p(t) = e^{-(at/d)^d}. \tag{4}$$

Here, a is a classic decay parameter, and d is a parameter determining the balance between early and later forgetting (Wickens, 1999). To this function may be added a parameter for initial learning, which we will label μ in order to correspond with the MCM, and a parameter b that determines the ultimate asymptote in

APPENDIX (Continued)

performance (Wickens, 1998). The resulting formula, for which we use the name *extended Weibull model* in this article, is

$$p(t) = b + (1 - b)\mu e^{-(at/d)^d}. \quad (5)$$

Testing Hypotheses

Initial learning. To test that initial level of performance is less than 1 in the Weibull model, one can test the full Weibull model as set out in Equation 5 against a submodel that has the μ parameter set at a default value of 1. The MCM has a performance level less than 1 at a retention interval of 0 as an automatic feature. No hypothesis on the initial level of learning can thus be tested in this model.

Asymptote in performance. Testing that the final asymptote in performance is above 0 in the Weibull model or the amended power curve requires testing the full model against one with the b parameter set at 0. In the MCM, a model with a positive asymptote in performance is one with the a_2 parameter set at 0 and μ_2 to a value larger than 0. This model can be tested against both a supermodel with a_2 larger than 0 and a submodel with μ_2 set at 0.

Shared retention function. Each model has submodels in which several curves share the decline parameter values. In the MCM, a_1, μ_2, and a_2 can be considered decline parameters, and a and d are decline parameters in the extended Weibull model. These can then be tested against supermodels in which decline parameters are allowed to vary for each curve.

Fitting the Data

These hypotheses are tested by fitting both sub- and supermodels with maximum likelihood estimation. The BIC for each model is then calculated by taking the natural logarithm of the likelihood and adding to it the factor $p * \log(n)$, in which p is the number of parameters of the model and n is the number of data points. Data are fitted here at the level of the individual data point, even though strictly speaking data points for 1 participant, or gathered with one question, are not independent. This adds to noise, but the large numbers of participants and questions ensure that the effects of this lack of independence are not large.

To see this, consider throwing an unbiased coin 50 times. The total number of heads will be distributed around 25, with most totals falling between 18 and 32. Now assume that we pick coins from a heap of very biased coins: Half of the coins on the heap yield 90% heads, and the other half yield only 10% heads. If we threw 25 times with one random coin from the heap and then picked another and threw that one 25 times, the expected number of heads would still be 25, but it could vary much more (if we picked 90% heads coins twice, we could expect 45 heads!). If we threw every biased coin just twice, however, and then picked another one from the heap, the variance of the number of heads would be only marginally larger than it would be with an unbiased coin.

(Manuscript received October 10, 2002;
revision accepted for publication September 18, 2004.)