

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

M. Meeter^{1,2}, J.M.J. Murre^{2,3} & S.M.J. Janssen²

Author note

Martijn Meeter, Department of Cognitive Psychology, Vrije Universiteit Amsterdam; Jaap M.J. Murre, Department of Psychology, University of Amsterdam; Steve M.J. Janssen, Department of Psychology, University of Amsterdam.

This study was supported by a PIONIER grant by the Dutch National Science Foundation (NWO) to the second author. We would like to thank Jeroen Raaijmakers for stimulating discussions.

Correspondence concerning this article should be addressed to Martijn Meeter, Department of Cognitive Psychology, Vrije Universiteit Amsterdam, VdBoechorstraat 1, 1081 BT Amsterdam, The Netherlands. Electronic mail may be sent to m@meeter.nl.

Abstract

A retention study is presented in which participants answered questions about news events, with retention interval varied within subjects between one day and two years. The study involved more than 14,000 participants and around 500,000 data points. The data were analyzed separately for those subjects answering questions in Dutch, and those answering questions in English, providing an opportunity for replication. We fitted models of varying complexity to the data, in this way testing several hypotheses concerning retention. Evidence was found for an asymptote in retention only in one data set. For participants with higher media exposure, a higher degree of learning was found but no difference in forgetting rate. Forgetting was thus independent of initial learning. Older adults were found to have similar forgetting curves as younger adults.

Introduction

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), while 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. It is, for example, 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 younger adults (Brainerd, Reyna, Howe, & Kingma, 1990; Cohen, Stanhope, & Conway, 1992; Wheeler, 2000). One reason for the long life of these controversies is disagreement on what would constitute a proper answer to the questions at hand; 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). Whether or not different conditions exhibit the same level of forgetting then 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: conclusion about parameter values are often bound to the model in which the parameters in question play a role (Rubin & Wenzel, 1996). An ideal study examining for example the dependence of forgetting on initial learning would therefore fit several models to the data, so that dependence or independence can be corroborated in different models.

Here, a study will be presented in which retention for news events is tested, involving around 14,000 participants and 500,000 data points. Participants were internet volunteers, who could log into an internet site after giving relevant personal details and take a test in which they answered questions about news events. An example of such a question was: “What was the name of the American country singer who died on September 12, 2003?” (q. 1430). Our primary goal in developing the internet site was to create a new retrograde amnesia test, by submitting news questions to web controls to test their appropriateness for inclusion in the test (Meeter, Murre, & Janssen, *subm.*). However, the control data is interesting in its own right, and can be used to study retention and forgetting. For each participant, thirty or forty questions were sampled concerning news events that had occurred at different moments in time. In this way, retention was measured at retention intervals ranging from a single day to up to two years.

The sheer size of the dataset in the study allow models to be fitted to the data and rejected with a high degree of precision. Several retention functions will be fitted to the data. Two functions are of particular interest: the Memory Chain Model (MCM) and what we will here refer to as the extended Weibull. Mathematical details of both models are given in the appendix.

The recently proposed Memory Chain Model (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 may be stored details that, if remembered, could trigger retrieval of the correct answer. A retrieval attempt is seen 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 function, 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 to it has an important advantage over others of the same function. Several features of retention functions have been parameterized in this formulation. For example, memory decline may not be constant over all retention intervals: forgetting is usually fast immediately after acquisition, decelerating at greater latencies (technically: hazard functions are 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: whether or not the model fits better with a certain parameter set to its optimal value, than with that parameter restricted to a default value (e.g., to 1).

A second aspects of retention that has been parameterized by Wickens (1998, 1999) is whether performance is perfect if the retention interval is extrapolated back to $t=0$. A third theoretically interesting aspect of retention that this function parameterizes is the final asymptote. Decline in retention may continue until finally performance is equal to zero, or performance may asymptote at a level above 0 or chance. Such an above-zero asymptote is what one would expect if there is some form of permastore in memory (Bahrick, 1984; Bahrick, Bahrick, & Wittlinger, 1975; Rubin, Hinton, & Wenzel, 1999). These aspects lead, together with a decline parameter, to a four-parameter function (see Table 1 and appendix).

Other retention functions often do not parameterize those aspects, 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 one, while there is no positive asymptote and forgetting is steeper at the early stages of retention than at later stages.

Table 1

Functions used in this paper to fit the retention data. For explanation of the formulas, see 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, or consolidation.

Amended power	$y = b + (1 - b)\mu(t + 1)^a$
Extended Weibull	$y = b + (1 - b)\mu^{-(at/d)^d}$
Memory Chain Model (MCM)	$y = 1 - \exp(-\mu_1 \{ e^{-a_1 t} + \frac{\mu_2}{a_2 - a_1} (e^{-a_2 t} - e^{-a_1 t}) \})$

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, than a steeper decline, followed by less pronounced forgetting (Chessa &

Murre, 2002). Whether or not the model exhibits an asymptote in forgetting is parameterized within it. In one form, it assumes consolidation from a first store to a second, more permanent one. If there is 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 to the question of whether decline parameters are shared by conditions such as different levels of initial learning, older adults versus younger adults, and recall versus recognition.

The aspects discussed above can be translated into three questions about retention, which, as already outlined, can in turn be translated into hypotheses about parameter values. These we tested in the two families of models, the Memory Chain Model (Chessa & Murre, 2002) and what we will call the extended Weibull function (Wickens, 1998, 1999). This allows the testing of different hypotheses by comparing models with more or fewer restrictions. By using two families of models with a very different mathematical structure, we aimed to untie our hypotheses from the specific model used. The appendix provides mathematical formulations of these 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 also used to investigate the issue of an asymptote in forgetting.

We tested two hypotheses related to the aspects discussed above (which model variants were used to test which hypothesis is described in the appendix).

1. Whether or not there is an asymptote in forgetting. In the extended Weibull function and in 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 a model without consolidation, and against one with both consolidation to and forgetting from store 2.

2. Whether or not different conditions can be fitted with certain shared parameter values. In particular, we tested whether decline parameters in both models remain constant under subject groupings with respect to age and initial learning, and for the two question formats (recall and recognition)

The web site contained both an international news test in the English language and a test aimed at the Netherlands in Dutch. As the questions and samples were different, the two data sets these tests generated will be discussed separately as Experiment 1 (Dutch test) and Experiment 2 (international test). To both data sets we will fit the functions discussed above. We will then look at how well models restricted by our hypotheses could fit the data. In experiments 1 and 2 data gathered up to February 21, 2003 was analyzed. Data gathered from then up to February 14, 2004 will be discussed as Experiment 3.

Experiment 1

Method

The internet news test, which we called Daily News Memory Test (DNMT), was part of a larger web site about memory aimed at the general public. On our general 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. It is still operative, but for experiment 1 we restricted our analyses to data gathered up to February 21, 2003. By then, the Dutch

version of the test had been completed 8244 times. The data sets discussed here and in the other two experiments in this paper can be found at www.neuromod.org/data.

Creation of the questions

Questions for the test were created through a tight script. Each working day, one of us [SJ] searched through a large daily newspaper and the website of television newscasts. Topics that had front-page attention in both were deemed suitable for questions. Moreover, only those stories that described dateable events were taken. A headline about this topic was then transformed into a question by taking one of the roles out and replacing it by an interrogative clause. This guaranteed that the questions 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 has won the 2000 presidential elections in the United States”, as its answer is contained in every news bulletin about Mr. Bush).

For the recognition version, three lures were created by freely associating on 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”. Participants were presented with the question and the four possible answers in random order, and were required to select a radio button before they could proceed to the next question. The recognition format is thus four-alternative forced choice (4-AFC).

Those questions to 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 4-AFC format, but participants were presented with a text field in which they could type an answer. Scoring of these answers occurred automatically by matching the answer of the participant against a word or part 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 enough prominence, but on others more than one question could be formulated. In all, 1006 Dutch questions were created in the 4 years analyzed in this paper.

Questions that proved too difficult were uninformative for the purpose of studying retention. We therefore checked both after thirty days and after sixty days whether participants surpassed chance performance on the question in the 4-AFC format. When performance was below chance (i.e., less than 25% of participants confronted with the question answered it correctly), the question was removed from the item list. This was the case for 92 Dutch items (not included in the total of 1006 questions), the data of which has not been included in the results presented here.

Design of the test

The questions were entered in 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 selecting questions from this database for presentation to the participant. Answers provided by the participants were stored in the same database.

At participants' first go, they had to register and answer some basic biographical questions. Our reasons for the elicitation of 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 participants took the test more than once; it is possible that a participant redid the test without logging in, but by registering under

a different name. As registering took several minutes and logging in just a few seconds, it is unlikely that this occurred often).

Once they were either registered or logged in, participants read a short instruction. Then they were presented with the questions. When the site went online, it contained only the Dutch test which then consisted of forty 4-AFC questions. From June 16, 2001 on, the format changed to ten questions in open format, and twenty questions in closed 4-AFC format.

The questions were not chosen entirely randomly. For several reasons questions were sampled more from recent time periods than from remote time periods. Questions were sampled without replacement one at a time at the moment that participants submitted their answer to the previous question to a database behind the site. With a likelihood of thirty percent, the new question was sampled from those created in the last thirty days. Another thirty percent of the questions were chosen from approximately the one-to-last month ([test day – 31] to [test day – 60]), while the remaining forty percent was older than that. A 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. The first way was through an inadvertent encounter with the web site while surfing on internet. Academic psychologists, for example, may have found the site through links on research sites. The second way was through a search engine. The site was indexed by several robots, and regularly turned up as an entry while searching for “memory” or “memory improvement”, as the internet site of which the DNMT was 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 possibility of sending friends an email with their score on the test and a challenge to improve on the participant.

By February 21, 2003, the Dutch version of the DNMT had been completed 8244 times by 4239 participants. Incomplete tests were discarded, as were the data of participants who registered with improbable dates of birth (implying an age of 100 years or of less than 5 years). In line with what is reported about the internet population (Buchanan & Smith, 1999), the participant group was disproportionately well-educated and tilted towards younger adults (see Figure 1 for a comparison with the general Dutch population). Sixty percent of subjects were male, while 40 percent were female.

Figure 1 about here

Results

Performance and performance predictors

In general, the DNMT proved to be moderately difficult. The average score was 42% correct on questions in open format, and 65% correct on the 4-AFC questions. Performance dropped with question age, with proportion correct at the longest intervals being around one half of initial performance for open questions, and around two thirds of initial performance for the 4-AFC questions.

Participants' average scores were analyzed as a function of the information that participants 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).

Table 2 about here

Regression of participant mean score to the biographical data provided by the participants. ‘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 who have the highest education, than it is for those who have the lowest education.

Predictors	Dutch		International-USA		International-other	
	norm. weights	difference over entire range	norm. weights	difference over entire range	norm. weights	difference over entire range
newspaper reading	0.28	0.13	0.13	0.05	0.16	0.07
gender	0.22	0.06	0.19	0.05	0.18	0.05
education	0.18	0.10	0.20	0.11	0.03	0.02
televised news	0.11	0.05	0.08	0.03	0.18	0.08
age	0.05	0.03	0.11	0.07	0.16	0.12

Analysis

In all, 238,547 data points were included in the analyses (data sets for all experiments can be found on www.neuromod.org/datasets). To ease fitting, we pooled data into bins of 3 days. Because 60% of the presented questions were less than two months old, there was a large drop in the number of observations for retention intervals longer than 60 days, as compared to those shorter than 60 days. Retention intervals longer than 60 days were therefore pooled into ten-day bins. Bins were labeled with their middle value (e.g., the three-to-five day bin was labeled as having a four day retention interval). For 4-AFC questions, the number of observations per bin was on average 3208, varying from 1163 to 9873. For open questions, the number of observations per bin varied from 585 to 4557 (mean: 1379). Figure 2 shows the resulting retention curves for the two item formats, 4-AFC and open questions.

As is customary in the literature, we have fitted curves that represent group averages. Conclusions reached on the basis of group data do not always apply at the level of individuals, and vice versa. If functions have parameters whose effects are nonlinear –as is the case for forgetting curves–averaging can yield functions that have a different form as 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 the Weibull function. Although this is a serious caveat, fitting curves to data of individuals would require more data per individual than was available here.

Figure 3 shows all model variants that were fitted to the data, organized in two separate hierarchies for the two model families that were considered, as well as the unrelated power model. Variants are ranked by the number of parameters they need to fit the two functions, from 8 to 3. Variants with less than 3 parameters produced such bad fits that they will not be mentioned. Each variant is connected to more and less restricted variants. A variant will be called a submodel of another variant if it is equivalent to the other variant in all aspects, except that one parameter has been set to a default value. The variant with fewer restrictions will be called the supermodel of the submodel.

Figure 2 about here

We fitted the models in Figure 3, with for the 4-AFC questions 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 with the more robust MSE method. The values reached by that method were taken as starting point for maximum likelihood estimation of parameter values. Of 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 hands. Following that, we fitted variants with intermediate numbers of parameters from two starting points: from the parameter values of both the supermodels and the 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 AIC) that its validity does not rest on the assumption that the true model is of the same family of models as the ones that are compared (Zucchini, 2000).

Figure 3 about here

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 2-store model with forgetting from the second store and shared forgetting parameters for recall and recognition [BIC 551.7, 5 parameters in total]. For the extended Weibull function family, it was a variant without asymptote or an initial learning parameter, and the parameter for the balance between early and late forgetting shared by recall and recognition [BIC 504.6, 3 parameters]. A power-law model with separate parameters for both curves also did well [BIC 545.6, 4 parameters]. Figure 3 reports the performance of all variants (reported as R^2 for higher intuitiveness), while Table 3 lists the parameter values of the best-fitting ones.

Only the retained MCM variant predicted an immediate performance (i.e., retention at an interval of 0 days) of lower than 1. This is a natural characteristic of the MCM. The retained Weibull model fitted better than a supermodel with immediate performance free to vary [BIC 511.5, 4 parameters]. The MCM predicted a performance of 79% correct for the 4-AFC 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; there is thus a maximum to performance –lower than 1– set by the number of participants who have acquired the news event in question. The retained Weibull variant, though predicting 100% correct on both 4-AFC and open questions, produced less than perfect performance at short intervals 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.

Table 3 about here

Parameter values of the best-fitting memory chain model (MCM) and extended Weibull function models for the whole data sets of the three experiments.

M.C.M.		μ_1	a_1	μ_2	a_2
Experiment 1	open questions	.92	.032	.018	.0010
	4-AFC questions	1.29	.032	.018	.0010
Experiment 2	open questions	.49	.018	.011	0
	4-AFC questions	.64	.018	.011	0
Experiment 3	open questions	.76	.020	.010	.00045
	4-AFC questions	1.20	.020	.010	.00045
Weibull		μ	a	d	b
Experiment 1	open questions	1	.0013	.20	0
	4-AFC questions	1	.00026	.20	.25
Experiment 2	open questions	1	.0089	.087	0
	4-AFC questions	1	.0011	.087	.25
Experiment 3	open questions	1	.0018	.17	0
	4-AFC questions	1	.00017	.17	.25

1. Final asymptote

Both retained models set the asymptote for retention at zero, and predicted negligible retention ($<10^{-5}$) if they were extrapolated to a retention interval of 10 years. Hence there was no evidence for an asymptote in forgetting. The MCM submodel with asymptote (2 stores, $a_2=0$) performed worse than the retained model, which had nonzero forgetting from store 2 [BIC 552.7, 4 parameters], while a Weibull supermodel with a nonzero asymptote parameter was also rejected [BIC 516.9, 4 parameters]. An amended power curve with asymptote also fitted worse [BIC 570.4, 6 parameters] than one without one [BIC 545.6, 4 parameter].

2. Shared parameters

Recall and recognition. In neither model family, correcting for guessing in the 4-AFC format was enough to fit the curves generated by the two question formats. The best-fitting MCM variant had separate learning parameters for the two formats, with shared decline parameters. A variant with separate decline parameters for the open and the 4-AFC curves was rejected [BIC 566.7, 8 parameters]. In contrast, the best-fitting Weibull variant had 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, 4 parameters].

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

In the fits above, it was assumed that with 4-AFC questions, participants who did not know an answer had a 25% likelihood of guessing the right alternative. It is possible that in fact subjects could 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 this would explain the differences in forgetting rate between open and 4-AFC questions, we compared the models described above to models in which all parameters were shared between the two formats, but an additional guessing parameter was introduced for the 4 AFC 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, 3 parameters], while in the MCM framework such a model fitted better than the model assuming differences in learning [BIC 533.9, 5 parameters]. Both models came to 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 vs. 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 4236 tests finished by participants who read many newspapers (6-7 per week) to the 2222 tests finished by participants who read newspapers sporadically (0-2 per week). Retention intervals were again pooled into 3-day bins for retention 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.

Figure 4 about here

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 partly obscure an asymptote in forgetting. However, the conclusion reported above of no asymptote was reached without a free guessing parameter (fits in the next experiment will also not put asymptote parameters against guessing parameters).

Table 4 about here

R² of the best memory chain model (MCM) and best extended Weibull function model to the different data sets fit in this paper. After each fit is listed the number of free parameters that were used to fit the data.

	all data		readers vs. nonreaders		older adults vs. college-age	
			nonshared forgetting	shared forgetting	nonshared forgetting	shared forgetting
Experiment 1						
MCM	0.974	5	0.971	10	0.969	7
Weibull	0.978	3	0.973	6	0.973	4
Experiment 2						
MCM	0.964	4			0.831	8
Weibull	0.960	3			0.821	6

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=1011.9, 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 the regular newspaper readers than for the sporadic newspaper readers in both the MCM and the Weibull model. The data could be accounted for even better in the MCM when the two groups were given separate guessing parameters in recognition [BIC 992.9, 7 parameter], suggesting that more knowledge allowed frequent newspaper readers to eliminate slightly more options in the 4-AFC format (guessing 41% correct, against 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 to participants between 18 and 24, which led to sample sizes of 1371 older and 1168 younger adults. Figure 5 shows the retention curves of both groups for the open and the 4-AFC items.

Figure 5 about here

Again, three-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 in both the case of the MCM [shared decline: BIC=879.3, 6 parameters; separate decline: BIC=842.6; 10 parameters], and in the case of 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 adults 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 4-AFC condition was allowed to vary between the two groups [BIC for the MCM: 817.8; BIC for the 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

To investigate whether the findings reported above were influenced by a lack of power, we studied what change to either the learning parameter or the main decline parameter from the optimal model would compensate in either the MCM or Weibull models for the loss of one parameter. It turned out that a 3% decrease or increase in the value of the learning parameter led to a reduction in the BIC equivalent of one parameter, just 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, and 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 a rejection of the model, indicating that the results obtained above were not due to a lack of power.

Theoretically, participants could answer 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 this would not have a large impact on our results, we performed two controls. First, we searched for individuals with a perfect score, which 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 the transportation times to and from the computer of the participants, they form a rough estimate of how many seconds a participant has spent on a particular item. There was no strong correlation between this reaction time and likelihood of a correct answer.

To formally test this we divided our data set into trials on which participants had spent more than 12 seconds (plus an estimated 3 for internet delays), and trials on which they spent 12 or less. There was no difference in the proportion of correct answers between fast and slow answers on the open questions, $t(6719) = .71$, $p < .48$. There was a slight difference on the 4-AFC questions, $t(7628) = 2.01$, $p = .045$. However, this was in the advantage of faster responses, with a mean 56.7% correct vs. 56.1% correct, contradicting the hypothesis that slow responses benefited on a significant scale from 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 the MCM framework suggested that recall and recognition have the same decline function, with the caveat that general knowledge aids recognition, so as to make the likelihood of guessing the right answer higher than mere chance. In support of this conclusion, it was found that groups that start out at a higher level of performance, consumers of the news and older adults, also have a higher guessing parameter than groups that start out at a lower level of performance: participants who read fewer newspapers and college-age adults. In both these comparisons, equivalent decline parameters were found for the two 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 4-AFC format. As the form of retention

was the same for the recall and 4-AFC format, however, it is unlikely that this can be interpreted as an asymptote only to be found in recognition.

In one instance the two model frameworks did not lead to the same conclusion, namely in the question of the initial level of performance. The MCM model has as a characteristic that retrieval is probabilistic, leading to suboptimal performance at lag 0 even when an item has been 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. This makes a theoretical initial performance of 100% extremely unlikely. In that respect the Weibull'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 Daily News Memory Test was opened to the public. It had the same form as the Dutch version. Questions were mostly translations from Dutch questions pertaining to international news, though some questions were taken from headings at internet sites dedicated to international news. In all, 418 usable English language questions were formulated. Another 50 were not used because participants scored less than chance performance in their multiple-choice format. Participants were asked to give the same information about themselves as in the Dutch test. As there is no widely known international system to code educational achievement, participants were asked how many years of formal education they had completed. In addition, participants were asked to list their country of residence.

Participants

The international version of the test was completed 9657 times by 7149 participants (only 19% of international participants performed the test more than once). About fifty percent of participants originated from the United States. Other mainly English-speaking countries were also well represented (e.g., United Kingdom, Canada and Australia, see Figure 6). As in our Dutch sample, international participants were comparatively young and well educated (see Figure 7). Of the international participants 46% was male, while 54% was female.

Figure 6 about here

Figure 7 about here

Results

Performance on the international test was lower than on the Dutch test, with participants answering on average 31% of the open questions correctly, and 52% of the 4-AFC questions. A regression analysis on participant means was done separately for participants with the U.S.A. as country of residence, and those originating from other countries (see Table 2). For American participants education and age were the best predictors of mean score, while for the remaining participants all variables except education were approximately equally powerful predictors (educational systems probably vary too much from country to country to let years of formal schooling be a good predictor). As there were differences between American participants and participants from other nations, we decided to only use data from American participants in our analyses of forgetting – this reduced a potentially large source of variability, and still left us with the majority of our data (5086 finished tests remained).

Analysis

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

Figure 8 about here

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

For a 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 a retention interval of 0) of lower than 1, while the retained Weibull model set immediate performance at 1. The MCM predicted a performance of 60% correct for the 4-AFC questions, and 39% correct for the open questions.

1. Final asymptote

A Weibull model with a nonzero asymptote parameter was rejected [BIC=501.2, 4 parameters], as was the case in the Dutch sample. The same was true for an amended power curve with an asymptote [BIC=518.4, 6 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, 5 parameters]. The best fitting MCM variant thus predicted an asymptote. Moreover, the best-fitting Weibull variant predicted such slow forgetting that performance was after very long intervals was not much worse than what the MCM predicted: performance on the open questions after 10 years was predicted to be 26.3% by the MCM, as compared to 17.7% correct predicted by the retained Weibull (and 21% by the amended power curve without asymptote).

2. 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, 6 parameters]. The retained Weibull variant had a separate decline parameter 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, 4 parameters]. This was similar to what was found in Experiment 1. Unlike in Experiment 1, however, assuming shared decline parameters but a likelihood greater than 0.25 of guessing the correct answer in 4-AFC did not provide an equally good description of the data. Such a variant had a worse fit in the MCM [BIC=507.2, 4 parameters], and in the Weibull framework [BIC=505.3, 4 parameters].

Participant age. As 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 in a similar way as in experiment 1. Using the same definitions as in experiment 1, we obtained samples of 251 older adults (age 60 and older) and 1346 younger adults (age between 18 and 24).

Figure 9 about here

Figure 9 shows the resulting retention curves. Again, we fitted these with the two retained models, and tested whether the data could be fitted while 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 in the MCM framework [shared decline: BIC=770.6, 6 parameters; separate decline: BIC=773.9; 10 parameters], and in the case of the Weibull function [shared decline: BIC=747.5, 4 parameters; separate decline: BIC=755.0; 6 parameters]. Both models set initial performance a little higher for the older adults than for the younger adults.

Discussion

In some respects, the results of experiment 1 and experiment 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 adults and younger adults showed the same decline function, even though the 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 pointed to one in the second experiment (the Weibull fits pointed to such slow decline that it is difficult to distinguish from an asymptote in forgetting). The second nonreplication was that 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. In experiment 2, this was not the case. This difference is convoluted with the issue of whether there is an asymptote in forgetting, as the guessing parameter functions as the asymptote in the retention curve associated with the 4-AFC format.

Experiment 3

To further investigate this issue of an asymptote in forgetting, we ran a third experiment in which retention was tested over a two-year period, instead of over a one-year period.

Method

Design of the test

In Experiment one and two, questions were sampled with a likelihood of 0.3 from the thirty days before the test, with a likelihood of 0.3 from approximately the one-to-last month ([test day – 31] to [test day – 60]), with the remaining forty percent being older than that. From February 28, 2003, the composition of the test was changed by sampling questions from five periods instead of from three. Fifty percent of questions were now sampled from the first two months, and the remaining questions were sampled from three periods: 30% from the period between 61 and 365 days before the test (i.e., questions were up to one year old), 10% from the period between 518 and 548 days before the test (i.e., questions were around 1.5 years old), and 10% from the periods between 700 and 730 (i.e., questions were around 2 years old). For the rest, the procedure was the same as in the previous experiments.

Participants

The new format was introduced for both the Dutch and the international version. However, as the number of international participants had by then experienced a dramatic drop, we restricted ourselves to analyzing retention for participants in the Dutch version (only 313 tests were finished by US participants, too few for reliable analyses). From February 28, 2003 to February 14, 2004, the Dutch version of the test was completed 3956 times by 2853 participants.

Results

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

Analysis

In all, 116,095 data points were included in the analyses. We again pooled data into 3-day and 10-day bins. This led to bins containing between 177 and 2682 data points. The resulting retention curves are shown in Figure 10.

Figure 10 about here

Fitting the curves with the MCM and the Weibull led to the same variants being retained as in experiment 1. For the MCM family this was a 2-store model with forgetting from the second store and shared forgetting parameters for recall and recognition [BIC 621.8, 5 parameters]. For the extended Weibull function family, it was a variant without asymptote

or an initial learning parameter, and the parameter for the balance between early and late forgetting shared by recall and recognition [BIC 627.0, 3 parameters].

Final asymptote

Neither of the retained models showed an asymptote in forgetting. Introducing an asymptote worsened the fit in both the Weibull model, in which it entails introduction of an extra parameter [BIC 636.7, 4 parameters], and in the MCM, in which it entails an extra restriction [BIC 628.1, 4 parameters].

To investigate the possibility that the asymptote is only evident in retention intervals longer than one 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 deduce from the fit of the whole two years, neither for the Weibull framework [BIC for the fit on the whole two year: 627.1, 3 parameters], nor for the MCM [BIC for the fit on the whole two year: 623.7]. 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- or overestimation of retention. This supports the conclusion reached in experiment 1 that there is no asymptote in retention on the Dutch version of the test.

Discussion

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

The best-fitting models pointed to a performance at a zero retention interval ranging from 74% for the Dutch 4-AFC questions (Experiment 1) to 46% for the international open questions (Experiment 2). These numbers may be equal to the proportion of the internet population that is exposed to the major headline items in the news. After initial exposure the proportion correct dropped to around one third of these values at the longest retention intervals (335 days) on open questions, and to around two thirds of initial performance on 4-AFC questions.

A feature of this research that needs some discussion –as it distinguishes the present retention study from previous ones– is the use of internet as a means to deliver the memory test to the participants. Though not many people would dispute that 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). However, what comparative research exists tends to show a general equivalence between data gathered via the internet and data gathered via traditional methods. For example, Buchanan and Smith (1999) found that a personality test on the internet taken by a sample of internet volunteers had the same psychometric characteristics as its paper-and-pencil pendant taken by a standard sample of psychology undergraduates. Even reaction time experiments delivered over the web elicited the same experimental effects as the same experiments in laboratory settings (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 as of the fact that 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. However, as retention interval was manipulated within participants and performance in absolute terms

was not important, possible sampling errors were not relevant in this study. In addition, possible disadvantages of research over the internet were more than outweighed by the possibility to include a very large number of participants in the study. In all, more than 14,000 participants took part in the test. This allowed for a wide range of questions surrounding retention and forgetting to be addressed in this study, and it gave the analyses a large power.

The fact that participants did not study the material-to-remember in a laboratory setting squarely places this research in the tradition of retention research with naturalistic retention material (Rubin & Wenzel, 1996). The materials used in this tradition have as a disadvantage that the study schedule is typically 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 to model relearning explicitly (not reported) 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, as 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, as 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.

1. 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 did predict a nonzero asymptote, while the extended Weibull function did not but predicted very slow forgetting (Experiment 2). The non-existence 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 re-normed every few years if indeed forgetting still took place after long intervals. Moreover, as 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 creation of questions section). With the exception of the study of Rubin et al. (1999), studies finding an asymptote have tended to rely on material such as school English or faces that have 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 be most likely to be overlearned. However, this is only speculative, as other aspects of our study may also explain our results. In particular, it is possible that a two-year interval is not enough to spot the emergence of an asymptote in retention.

2. Shared parameters

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

Recall and recognition

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

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

Less steep forgetting for the 4-AFC 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 did 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 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 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 findings of several other studies (Rubin & Wenzel, 1996).

Aging

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

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

intervals have varied over a relatively wide range tend to find stronger effects of aging on 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, as 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 younger adults were less educated than in typical studies (not all at university level), and older adults 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 may lead to interpretational problems (Loftus, 1985) that were thus avoided in the current study.

Perhaps the most likely explanation concerns the material used. In many studies the material consists of lists of words or pictures acquired in one experimental session. Here, it consists 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 (see Einstein, McDaniel, Manzi, Cochran, & Baker, 2000 for further evidence).

Conclusion

Rubin and Wenzel (1996) state that comparisons of parameter values in different conditions depend on models in which the parameters feature. The example given is 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 the data using two, and sometimes three models. By and large, where a hypothesis was tested using different models, the same conclusion was reached.

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 and 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, and 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*. Amsterdam: NeuroMod Technical Report 02-01.
- 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*. [About 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 and 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, and 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. (subm.). 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 and 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, and 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 and Cognition*, 9, 384-397.
- Wheeler, M. A. (2000). A comparison of forgetting rates in older and younger adults. *Aging, Neuropsychology, and Cognition*, 7, 179-193.
- Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory and 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 reflection on the 30th anniversary of the Atkinson-Shiffrin model*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44, 41-61.

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 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 time t (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 paper 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 [$p(\text{correct})$]: the power function goes to infinity at values close to zero, whereas $p(\text{correct})$ can only vary between 0 and 1. With a slight change it starts at 1 and still is 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 $p(\text{correct})$ is the relationship between this measure and the underlying memory strength. A model in which this relationship has been explicitly modeled is the Memory Chain Model (Chessa & Murre, 2002).

Chessa and Murre (2002) propose 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: This leads 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 a decline parameter a :

$$\text{Equation 1: } r(t) = \mu e^{-at}$$

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. From sensory registers memories may be copied to STM, from there to a hippocampal long-term memory, and from there into a neocortical memory. In most fits, Chessa and Murre use 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 in store 2 with a constant likelihood of μ_2 . From this store, they are lost with a likelihood a_2 . This results in the following function for the intensity (expected number of memories) at any time point t .

$$\text{Equation 2: } r(t) = \mu_1 \left\{ e^{-a_1 t} + \frac{\mu_2}{a_2 - a_1} (e^{-a_2 t} - e^{-a_1 t}) \right\}$$

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, by the Poisson distribution, to the natural exponent of minus the intensity at time point t (expected number of memories). The resulting retention function is:

$$\text{Equation 3: } p(t) = 1 - \exp\left(-\mu_1 \left\{ e^{-a_1 t} + \frac{\mu_2}{a_2 - a_1} (e^{-a_2 t} - e^{-a_1 t}) \right\}\right)$$

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:

$$\text{Equation 4: } p(t) = e^{-(at/d)^d}$$

Here, a is a classic decay parameter, while 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 for correspondence with the MCM, and a parameter b that determines the ultimate asymptote in performance (Wickens, 1998). The resulting formula, for which the name ‘‘Extended Weibull Model’’ is used in this paper, is:

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

Testing hypotheses

Initial learning. To test that initial level of performance is lower 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 below 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. To test that the final asymptote in performance is above 0 in the Weibull model or the amended power curve, one tests the full model against one that has the \underline{b} parameter set at 0. In the MCM, a model with a positive asymptote in performance is one with the \underline{a}_2 parameter set to 0, and μ_2 to a value larger than 0. This model can be tested against both a supermodel with \underline{a}_2 larger than 0, or a submodel with μ_2 set to 0.

Shared retention function. Of each model there are submodels in which several curves share the decline parameter values. In the MCM, \underline{a}_1 , μ_2 , and \underline{a}_2 can be considered decline parameters, while \underline{a} and \underline{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.

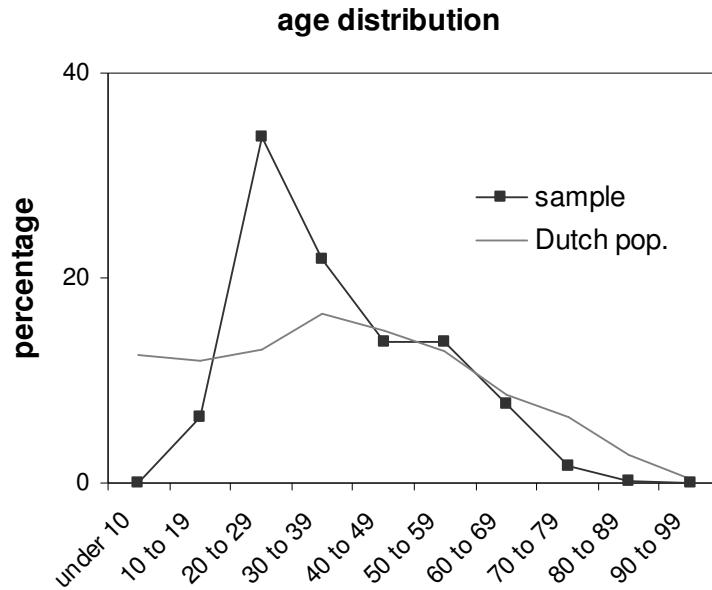
Fitting the data

The hypotheses are tested by fitting both sub- and supermodels with maximum likelihood-estimation. The Bayesian Information Criterion (BIC) of each model is then calculated by taking the natural logarithm of the likelihood, and adding to it a factor $p * \log(n)$, where p is the number of parameters of the model, and n the number of data points. Data is fitted here at the level of the individual data point, even though strictly speaking data points from one subject, or gathered with one question, are not independent. This adds to noise, but the large numbers of subjects and questions ensures that the effects of both are not large.

To see this, consider throwing an unbiased coin fifty times. The sum of the number of heads will be distributed around 25, with most sums falling between 18 and 32. Now assume we pick coins from a large heap of very biased coins: half of the coins on the heap yield 90% head, the other half 10% heads. If we throw 25 times with one random coin from the heap, then pick another and throw that one 25 times, the expected number of heads is still 25. However, the number of heads can vary much more (if we twice picked a 90% heads coin, we can expect 45 heads!). If we used every biased coin just twice, however, and then picked another one from the heap, the variance of the sum of heads is only marginally larger than with an unbiased coin.

Figure 1 Distribution of (a) age and (b) education in the Dutch sample. Both are compared to the distribution of both variables in the general population of the Netherlands ('Dutch pop.'; source: www.cbs.nl). Education was, in accordance with statistical convention in the Netherlands, classified by highest attained educational grade.

a.



b.

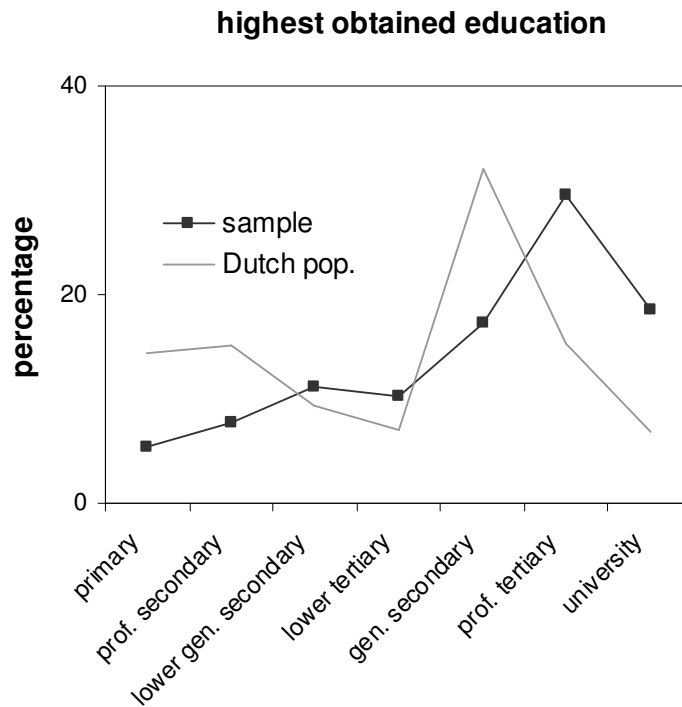


Figure 2 Retention curves for the Dutch sample (Experiment 1). Plotted separately are the data from the open questions and 4-AFC questions. Continuous lines represent the best-fitting MCM variant with parameters listed in Table 3.

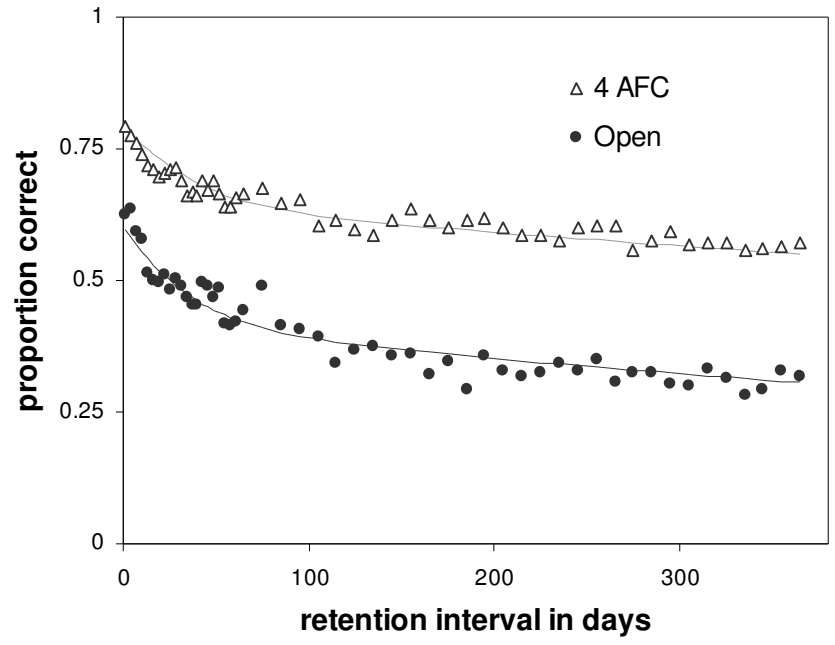


Figure 3 Comparison of all models when simultaneously fit on the open-question and 4-AFC question data for the Dutch sample (Experiment 1). Models are ranked according to their number of free parameters, while goodness of fit is color-coded (fits also between brackets behind the model name). Supermodels are connected with their submodels through arrows. Models with a black cadre were retained for further consideration; see text. 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.

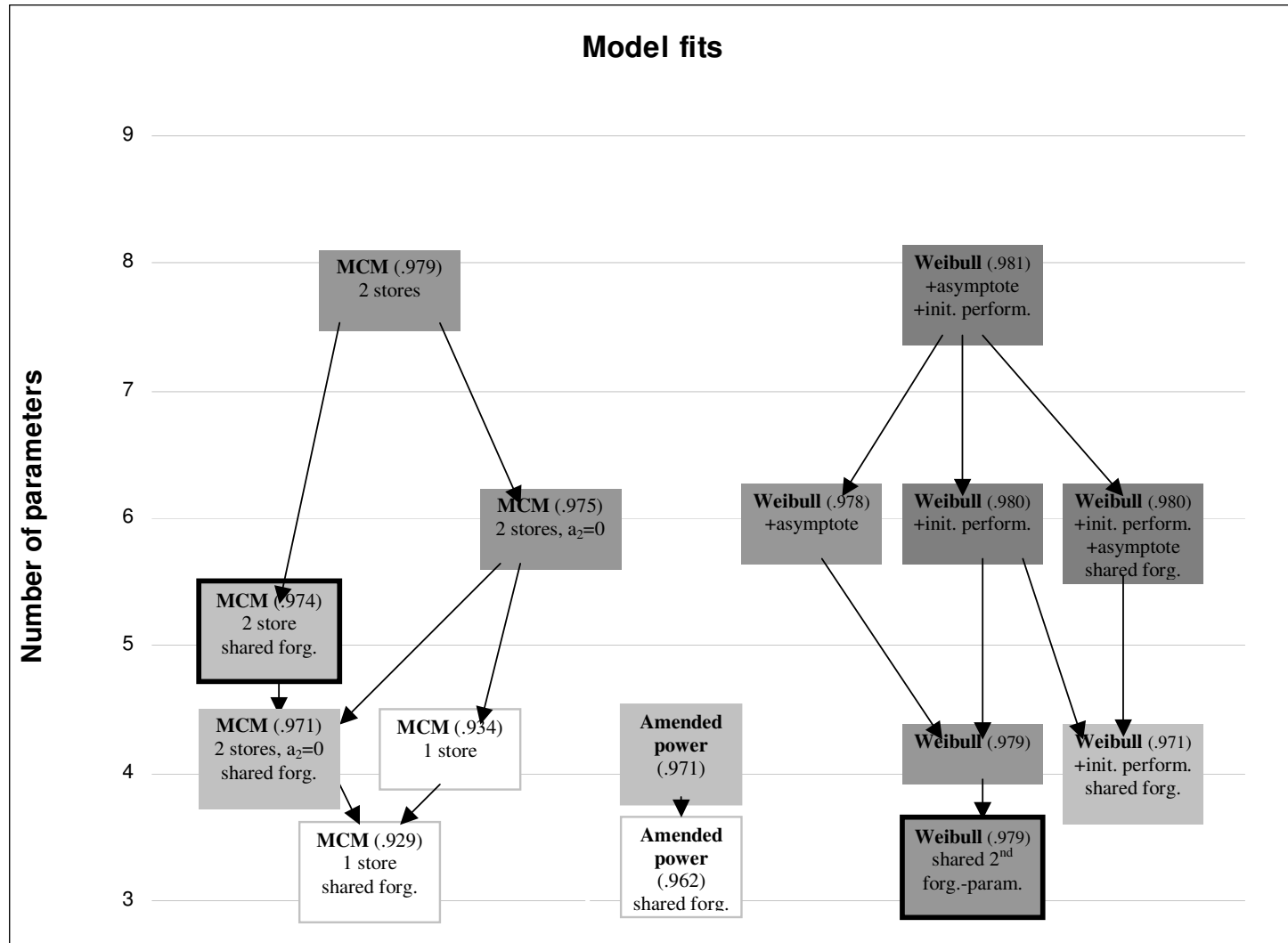


Figure 4 Open (a) and 4-AFC (b) retention curves for the Dutch sample (Experiment 1). Plotted separately are those participants who read many newspapers (at least 6 a week) and those who read few newspapers (0-2 a week).

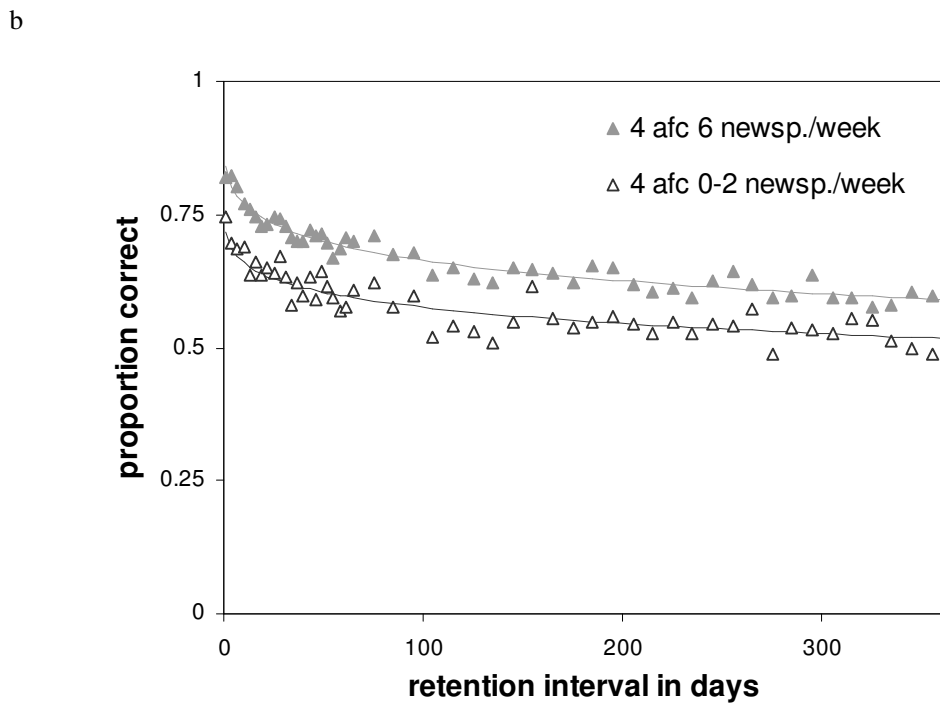
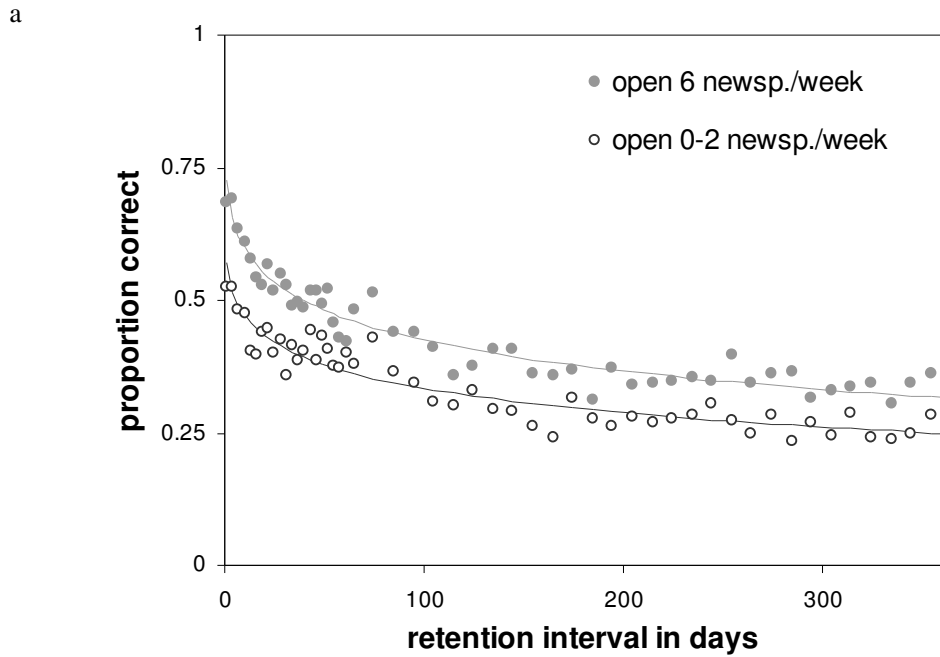
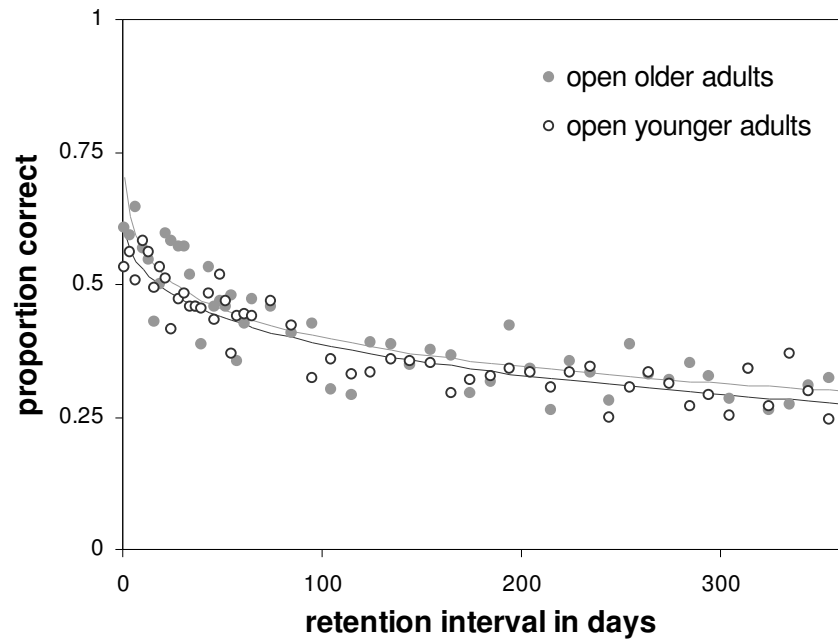


Figure 5 Open (a) and 4-AFC (b) retention curves for the Dutch sample (Experiment 1). Plotted separately are older (aged 60 or more) and college-age participants (age 18 to 26).

a



b

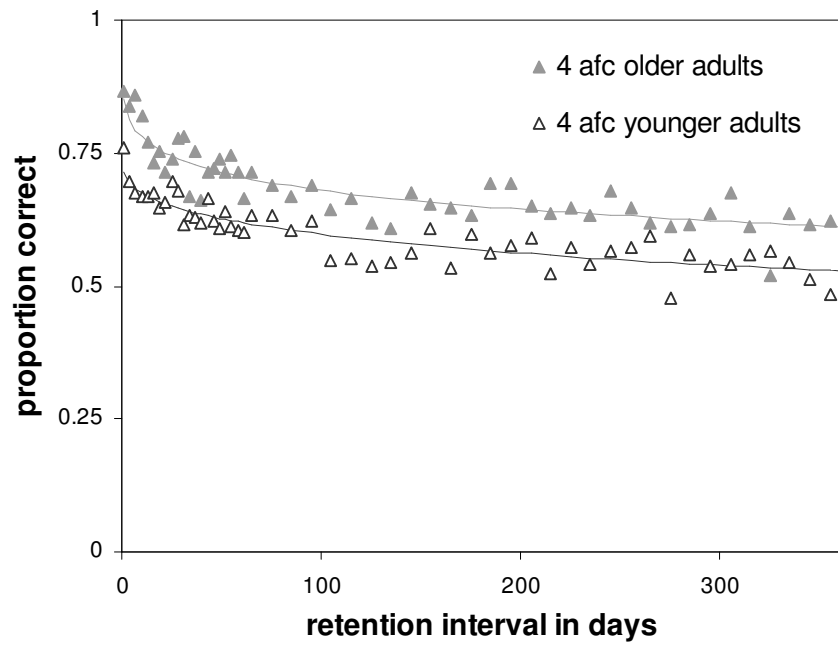


Figure 6
(Experiment 2).

Country of residence of participants in the international sample

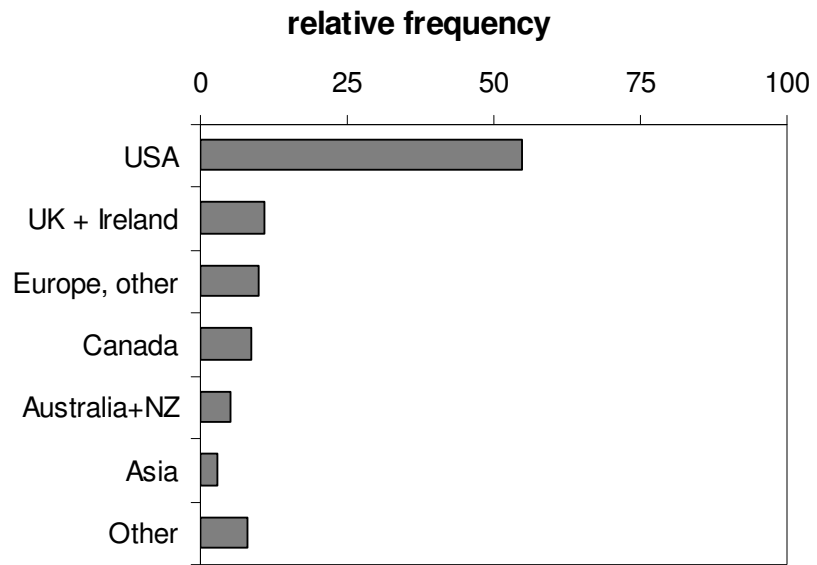
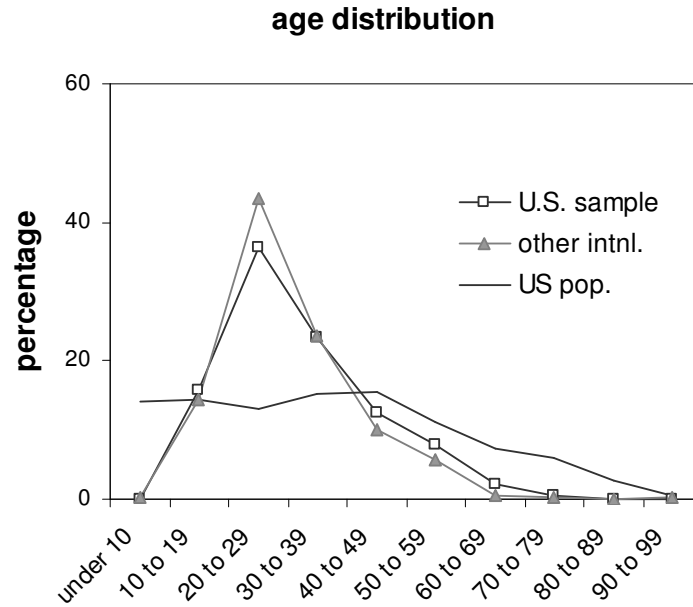


Figure 7 Distribution of (a) age and (b) education in the international sample (experiment 2). Participants originating from the USA are plotted separately from those originating from other countries. For age, an estimate of US population distribution was added (source: www.census.gov). Education was elicited as the number of finished years of formal education.

a.



b.

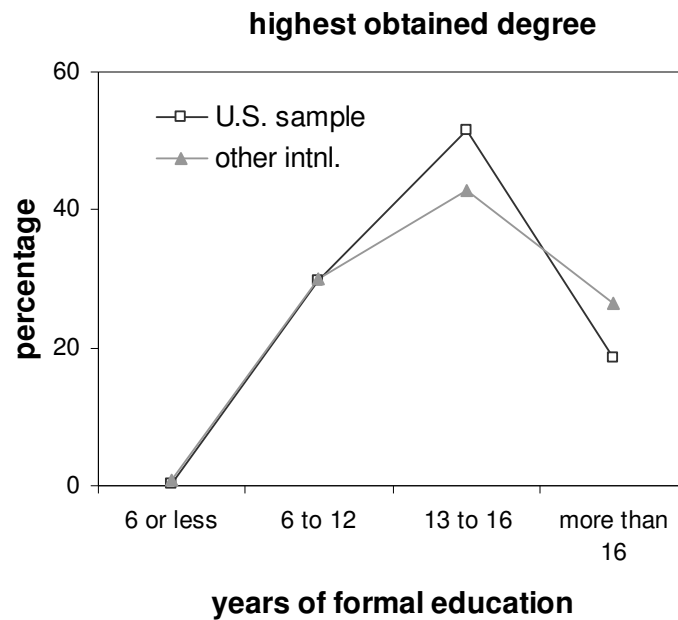


Figure 8 Retention curves for the international sample (Experiment 2). Plotted separately are the data from the open questions and 4-AFC questions

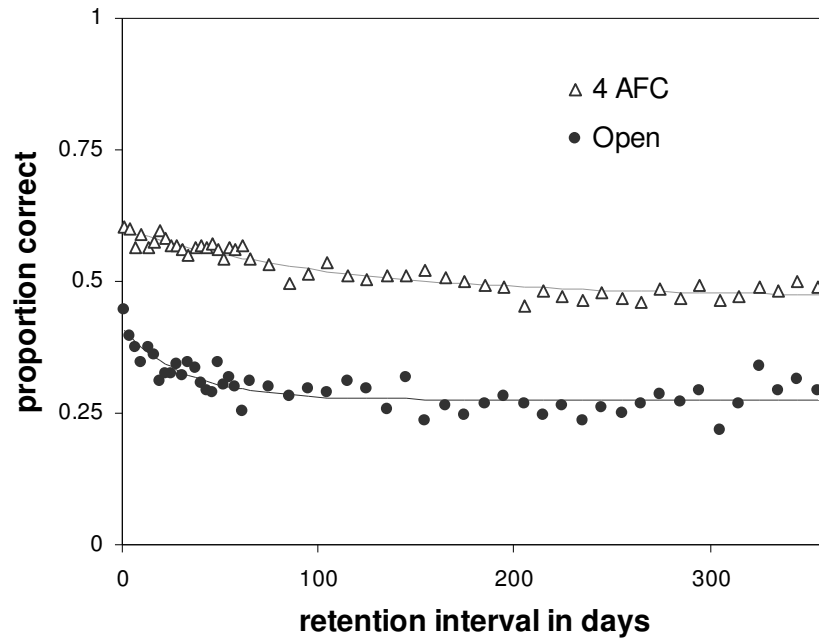
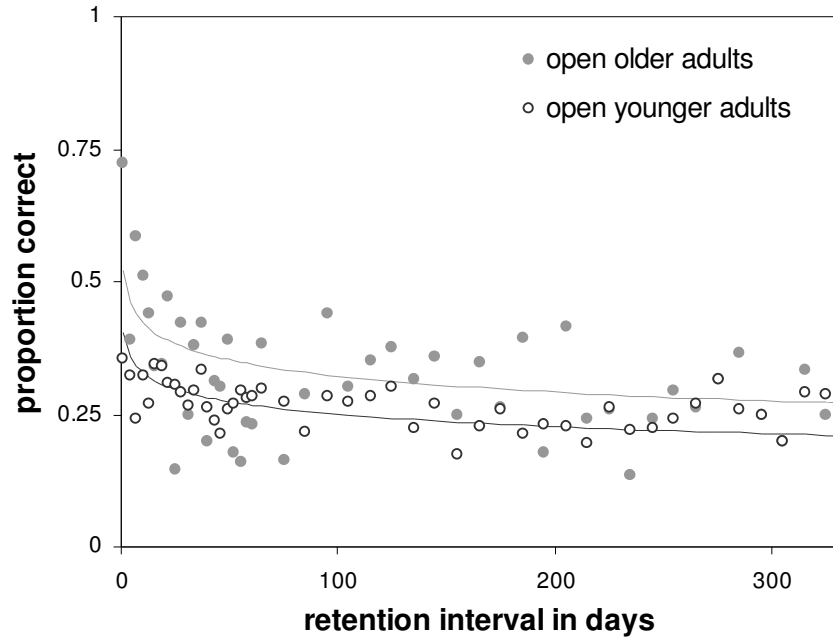


Figure 9 Open (a) and 4-AFC (b) retention curves for the American participants in the international sample (Experiment 2). Plotted separately are older (aged 60 or more) and college-age participants (age 18 to 26).

a



b

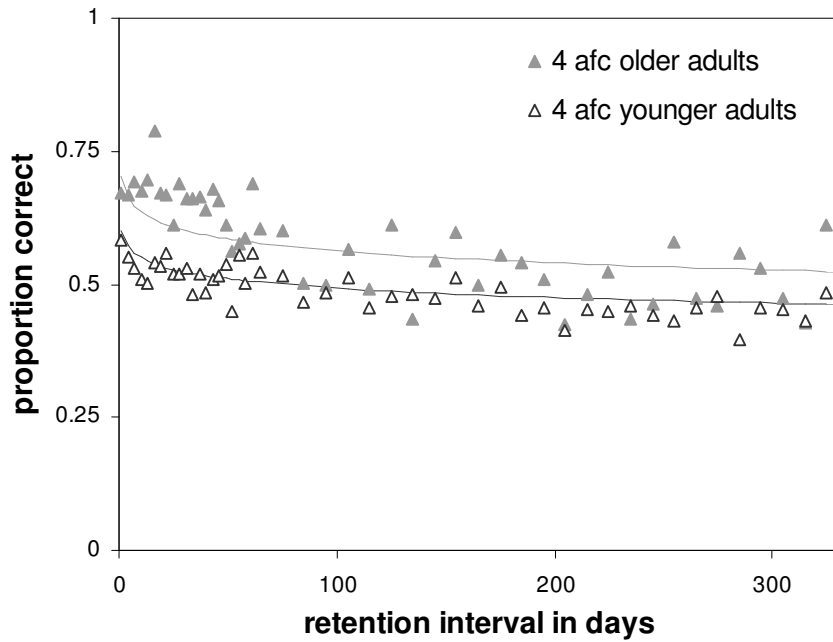


Figure 10 Two-year retention curves for open (a) and 4-AFC (b) questions for the Dutch sample (Experiment 3). Continuous lines correspond to the fits of a two-store MCM with decline parameters shared by open and 4-AFC questions, with parameter values fitted only on the first year of retention.

