Running head: POWER FORGETTING

The Form of the forgetting curve and the Fate of memories

Lee Averell and Andrew Heathcote

The University of Newcastle

Abstract

Psychologists have debated the form of the forgetting curve for over a century. We focus on resolving three problems that have blocked a clear answer on this issue. First, we analyzed data from a longitudinal experiment measuring cued recall and stem completion from 1 minute to 28 days after study, with more observations per interval per participant than in previous studies. Second, we analyzed the data using hierarchical models, avoiding distortions due to averaging over participants. Third, we implemented the models in a Bayesian framework, enabling our analysis to account for the ability of candidate forgetting functions to imitate each other. An exponential function provided the best fit to individual participant data collected under both explicit and implicit retrieval instructions, but Bayesian model selection favored a power function. All analysis supported above chance asymptotic retention, suggesting that, despite quite brief study, storage of some memories was effectively permanent.

KEYWORDS: Forgetting, hierarchal models, Bayesian model selection.

The Form of the forgetting curve and the Fate of memories

The Form of the forgetting curve and the Fate of memories

The search for a general quantitative description of the "forgetting curve", the nonlinear function relating the observed probability of memory retention (R) and the delay or lag between study and test (t), is one of experimental psychology's oldest problems (Ebbinghaus, 1885/1974). This problem has been raised for both short-term and long-term memory (Wickens, 1998); here we focus on the latter. Although the form of the forgetting curve is still seen as being of "central theoretical importance"(G. D. A. Brown, Neath, & Chater, 2007) over a century of research has failed to result in a consensus. The lack of consensus after so much effort has led some to question the utility of the entire enterprise of attempting to identify general laws for memory (Roediger, 2008).

In this paper we attempt to determine the form of the forgetting curve using *hierarchical models* and *Bayesian model selection* (see Shiffrin, Lee, Wagenmakers, & Kim, 2008 for a tutorial). These methods address two potential problems with previous analyses, distortions in group analyses based on retention data averaged over participants, and differences between candidate forgetting functions in complexity, which determines their ability to imitate each other in noisy data (Myung & Pitt, 1997). Lee (2004) found that complexity varied substantially among the set of five two-parameter retention functions identified by Rubin and Wenzel (1996) as providing the best fit[1] to 210 published forgetting curve data sets. Due to these differences, his Bayesian analysis, which penalized more flexible models as a function the level of measurement noise, found cases in which functions that gave a worse fit had a higher probability of being the true model than functions which gave a better fit. These results imply that inconstancy in results about the form of the forgetting curve might arise because of the combined effect of variations in the flexibility of the candidate functions and in the level of measurement noise between different studies.

Averaging can be problematic for indentifying the form of the forgetting function

when participant curves vary in shape. That is, when participant curves are not related by a linear transformation. In such cases the shape of the average curve can be different from that of any individual curve, with the degree of departure depending on the amount of shape variation amongst participants (S. Brown & Heathcote, 2003b). Hence, inconsistency in findings about average forgetting curves may simply reflect incidental variations in the degree of individual differences in forgetting curve shape between studies. Although individual curve analysis avoids the averaging distortion, it can be plagued by high levels of measurement noise, which can also lead to inconsistent results and exaggerate confounding due to complexity differences. Indeed Cohen, Sandborn, and Shiffrin (2008) showed that in simulated experiments with few data points per individual, and hence high levels of measurement noise, the probability of selecting the data-generating forgetting function was better for group than individual analysis. Hierarchical models, which do not average data but have the advantages of group level analysis in terms of reduced measurement noise, offer a potential solution to this dilemma. The reduction in measurement noise as a result of the hierarchical structure is due to the pooling (shrinkage) of individual participant parameter estimates around a common mean.

In the next section we introduce a general framework that identifies two components to the question about the form of the forgetting function. What mathematical function characterises the nonlinear change in retention with lag? Do forgetting curves have an asymptote (a) greater than chance performance (g)? We will refer to these respectively as the "function" and "fate" questions. We then attempt to answer both questions by analysing a cued recall data set collected by Averell and Heathcote (2009). In their experiment, participants studied 4-6 letter words and at test were cued with a stem consisting of the first three letters of a studied word. In one condition participants were given explicit memory instructions; they were asked to complete the stem to make a studied word. In a second condition a different group of participants were given implicit

memory instructions; they were asked to complete the stem with the first word that came to mind. Retention was measured at seven lags ranging from around one minute to one hour in the first experimental session, and again in sessions that occurred 1, 7 and 28 days after the initial session.

The initial session was modelled after a similar experiment performed by McBride and Dosher (1997). They found constant retention for lags greater than 15 minutes, but suggested that a "further decline would be measured in hours or days" (p. 380). Averell and Heathcote (2009) included the last three sessions in order to test this possibility, and to provide data that strongly constrained the answer to the fate question. In order to reduce measurement noise, so the data also strongly constrains the answer to the function question, each participant responded to a large number of tests at each lag, around 80 in the first session and 104 in later sessions.

Candidate forgetting curve forms

Equation 1 is a general expression for the forgetting curve.

$$R(t) = a + (1 - a) \times b \times P(t) \tag{1}$$

P varies nonlinearly with t as a function of θ, a vector of positive parameters. We assume that, for all $\theta, P(0) = 1$ and that $P(t)$ approaches zero for large values of t. The parameters a and b are also assumed bounded between zero and one, and hence $R(t)$ is similarly bounded, which must necessarily be the case as $R(t)$ is a probability. Enforcing this bound is important as otherwise data fits can be inflated (see Navarro, Pitt, & Myung, 2004 for further discussion). Values of b less than one allow for the possibility that $R(0) < 1$, which might occur, for example, if study encoding fails [2]

In terms of Equation 1, the function question is answered by identifying $P(t)$ and the fate question is answered by determining if $a > g$. In cases where retention is measured by responses chosen from a very large set (e.g., cued recall of unrelated word pairs) it can

be assumed that $g = 0$. However, in Averell and Heathcote's (2009) experiment each test stem could only be completed by a relatively small set of words (four or more), so chance performance had to be taken into account. An initial calibration study determined that $g = .116$ for their stimuli, so we estimated a parameter $\hat{a}$ that was bounded between zero and one and that was related to the asymptote by $a = .116 + (1 - .116) \times \acute{a}$.

Opinions are strongly divided on the question of the fate of memories. Chechile (2006) stated that "The inability of a function to account for the possibility of permanent retention is a serious failing" (p. 36). In contrast, Wixted (2004a) asserted that a chance asymptote "seems to be the view of almost everyone who has ever investigated the mathematical form of forgetting" (p. 871). Wixted demonstrated that the fate and function questions are intimately connected. For example, an exponential function provided a much worse fit than a power function when fit to free recall data reported by Wixted and Ebbesen (1991) when both had no asymptote, but fit equally well when both had an asymptote.

We considered three candidate forms for the function P, an exponential function with parameter $\alpha, P = e^{-\alpha}$ where α represents the rate of forgetting. A Pareto function with parameters γ and β, $P = (1 + \gamma t)^{-\beta}$ where γ scales the effect of β, the rate of forgetting (see below). Lastly, a special case of the Pareto, a power function, in which it is assumed that γ=1. The additive constant in the latter two functions ensures that $P(0) = 1$, and its value is fixed at unity without loss of generality when the b parameter is also estimated, as for any other value $k, b(k + \gamma t)^{-\beta} = \acute{b}(1 + \acute{\gamma} t)^{-\beta}$, where $\acute{b} = bk^{-\beta}$ and $\acute{\gamma} = \gamma/k$. The same argument shows that the hyperbolic function examined by Rubin and Wenzel (1996), and favoured by Lee's (2004) Bayesian analysis, $1/(mt + b)$, is a special case of the Pareto where $\beta = 1$. Although this set of functions is not exhaustive, it does cover most of the best plausible candidates from previous studies[3], and we contend that it also captures important characteristics of the psychological mechanisms thought to be

responsible for the form of the forgetting function.

The relationship between our candidates, and their psychological interpretation, is illustrated comparing shapes as measured by their hazard function $H(t) = (-dP(t)/dt)/P(t)$ (Chechile, 2006). For the exponential, the hazard function is a constant, $H(t) = \alpha$, and for the Pareto it is a hyperbolically decreasing function of lag, $H(t) = \gamma\beta/(1 + \gamma t)$. The hazard function shows that, for the exponential, the rate at which memories are forgotten is a constant proportion of the remaining memories which can be forgotten. For the Pareto and power function, in contrast, something is slowing down the rate of forgetting relative to the exponential as lag increases. Wixted (2004a) attributed the slowing to consolidation, a process that makes memories less vulnerable to forgetting as they age. He related the candidate forgetting functions to Jost (1897) second law of memory, which states that if two memories have an equal strength at lag t, forgetting will be more rapid for a younger memory than an older memory thereafter. Both Pareto and power functions are consistent with Jost's law, whereas the exponential function is not, as once their strengths are equal, both older and younger memories must be forgotten at the same rate if forgetting is exponential (Simmon, 1966).

Pareto and power functions differ only in the scale on which consolidation occurs. For example, if $\gamma = 0.1$ in the Pareto function, the effect of an increase in t is ten times less than for a power function. Hence, for small values of γ consolidation is slow, whereas for large values of γ it is fast. Wixted's (2004a) analysis demonstrates that weak consolidation might be mistaken for an asymptote, as it results in very gradual rate of decrease at longer lags. For example, in fits to Rubin, Hinton, and Wenzel's (1999) data on cued recall of unrelated word pairs, he found that an exponential function provided an accurate fit with an asymptote of .11, whereas a Pareto function with an asymptote fixed at zero provided a slightly better fit with $\gamma = 0.11$.

In light of such findings, and related findings favoring the zero-asymptote Pareto

with a range of estimated values, Wixted (2004a) contended that that the fate of memories that are not rehearsed after initial study, even memories that are initially very strong, such as in Bahrick's (1987) study of high school knowledge of Spanish, are eventually completely forgotten. That is, although consolidation slows the rate of forgetting, it is ultimately ineffective. The implication is that forgetting functions should not include an asymptote parameter, but that they should allow for consolidation to occur on a range of different time scales. These implications are captured by the Pareto function with a zero asymptote.

In light of these considerations, we fixed the asymptote of our candidate Pareto function at chance performance in our analysis. That is we assumed:

$$R(t) = .116 + (1 - .116) \times b \times (1 + \gamma t)^{-\beta} \tag{2}$$

For the power and exponential functions, in contrast, we estimated the asymptote (taking into account chance performance as previously discussed), to allow, respectively, for ultimately effective consolidation and no consolidation. The inclusion of an asymptote parameter in the power function shows that while the power model is a special case of the Pareto it is not nested within it.

$$R(t) = a + (1 - a) \times b \times (1 + t)^{-\beta} \tag{3}$$

$$R(t) = a + (1 - a) \times be^{-\alpha t} \tag{4}$$

Henceforth, we will refer to these candidate functions, each of which has three estimated parameters, simply as the Pareto, power and exponential functions. Comparison of all three bears on the function question, whereas comparison of the Pareto with the other two bears on the fate question [4]

Data Analysis

Complete details of experimental methods are given in Averell and Heathcote (2009); here we provide an overview that highlights aspects that are important for answering the questions at hand. The 32 participants (half in the implicit and half in the explicit condition) performed thirty 4.3 minute study-test cycles in the first session, with an 8.6 minute break between the 16^{th} and 17^{th} cycles. Otherwise breaks between study and test, and between study-test cycles, were only 7 second to ensure that participants had little time for rehearsal of the study words. Study consisted of 17 word pairs being presented for 4 seconds each, with participants required to rate which word occurred more frequently in their linguistic experience. At test 26 stems were presented sequentially for 7 seconds each, and during each presentation participants were required to type a completion. In later sessions, which were performed in the same room, the same procedure applied, except that five study-test cycles were performed with no long break, and only 13 pairs were studied on each cycle. The first cycle in later sessions was a warm up, whereas in the remaining cycles test stems corresponding to words studied in the first session. For each participant no study word or test stem was ever repeated in the entire experiment.

Several aspects of the experimental methods bear on two important and related issues, retrieval failure and interference. Retrieval failure occurs when an available memory (i.e., one that is still in storage) is not accessible at the time of testing. Such failures can occur when memory is probed with retrieval cues that are not strongly associated to the target memory, or due to interference occurring when other memories out-compete with the target memory for retrieval. If the level of retrieval failure differs across lags any answer to the function question would be confounded, as the shape of the forgetting curve would be altered by the differences. Strong retrieval failure at long lags would also confound the answer to the fate question, as available memories might not result in above chance performance.

Averell and Heathcote (2009) used stem-cued testing with the aim of minimizing both effects. Only one word consistent with each test stem was studied, which should minimize interference due to retrieval competition effects, because no allowable non-target test response was ever studied. Stems provide strong retrieval cues, so particularly in later experimental sessions, when cues related to the study context are less likely to be used, stored memory traces are more likely to still be accessible. To further reduce the possibility of retrieval failure in later sessions, each participant videoed a first-person view of their walk into the experimental room from the foyer where they were met by the experimenter. The experimenter also made a short video of the participant sitting in front of the experimental computer in order to capture aspects of the study context that might not be present in later sessions (e.g., the participant's attire). Prior to the experiment participants also answered questions about the weather, their surrounds and mood as well as their activities just prior to commencement of the first session. The answers to these questions and the videos were reviewed just prior to the commencement of later testing sessions.

Three further measures were taken to also reduce confounding by factors that differed between lags. The number of stems in a test cycle that corresponded to words studied in the same test cycle was approximately equated over all ten lags. This control aimed to equate the degree to which recall of one item could assist recall of following test items from nearby study positions (Howard & Kahana, 2002). Testing of items in the shortest (1.2 minute) lag condition occurred around one quarter of the way through the test cycle following the cycle in which they were studied. The intervening test trials made it unlikely that rehearsal for this lag would advantage performance relative to longer lags. The remaining lags occurred three quarters of the way through the test list and 1, 2, 4, 8 or 16 cycles later, on average lags of 2.93, 6.45, 10.75, 19.35, 36.55 and 70.95 minutes. The lags for the following three sessions were, on average, 1,440, 10,080 and 40,320 minutes

after study. An increasing spacing was used in order to provide the densest measurements of the forgetting curve in the region where it was most rapidly changing (Myung & Pitt, 2009).

Finally, the seven lags in the first session were dispersed over the first 2.1hour experimental session so that their average midpoints were close to equivalent (69.8, 70.4, 70.2, 70.4, 70.2, 70.1 and 70.1 minutes into the session). This equivalence minimized the possibility that that the lag effect within the first session was confounded by fatigue or differential interference effects related to the position of the lag in the test session, whether specific to a test item or non-specific. However, it is important to note that this final control does not apply to the three longest lags. For example, the later testing sessions took only 35 minutes, and so performance may have been improved by a reduction in fatigue. On the other hand performance in the later sessions may have been reduced by a build up of retroactive interference after the first session or because, despite the measures taken, the reinstatement of study context in later sessions was not equivalent to the first session. In light of these possibilities, after reporting results for the analysis of all lags, we discuss parallel results obtained based on only the lags in the first session.

Maximum Likelihood Analysis

Individual and group analyses characterize each individual's data, or the group average, by estimating a set of retention function parameters. Retention data, in the form of counts for correct responses at each lag (n_i for $i = 1 \ldots$ T *lags*) is usually modelled by a binomial distribution, $n \sim B(p_i, N_i)$ where the N_i are the number of responses at each lag and the p_i are the probabilities of a correct response at each lag. The binomial probability parameters, in turn, are assumed to come from a retention function, such as Equations 2-4 with parameter vectors of the form $\Theta(a, b, \theta)$. Estimation of this type can be done by the method of maximum likelihood, using an optimization algorithm to find an estimate, $\hat{\Theta}$,

that minimizes the deviance, which equals -2 times the log-likelihood. The minimum deviance, D, is obtained by plugging $\hat{\Theta}$ into its retention function to obtain retention probability estimates, $\hat{p}$ which are in turn substituted into the following equation:

$$D = -2\arg\max \sum_{i=1}^{t} n_i \ln p_i + (N_i - n_i) \ln(1 - p_i) \tag{5}$$

A summary of the group results can be made by summing of the individual deviances, as each deviance, and so their sum, have a χ^2 distribution. When we performed this analysis the exponential function clearly had the best fit with total deviance values of (959 and 874) for the explicit and implicit data respectively, with the power function being intermediate (1032 and 902) and the Pareto function providing the worst fit (1070 and 921).

At the individual participant level 11 of the 16 participants in the explicit instruction condition had a lower deviance for the exponential compared to the power and 14 of the 16 participants had lower deviance for the exponential when compared to the Pareto. In the implicit instruction condition 12 of the 16 participants have lower exponential deviance relative to the power model while 13 of the participants had lower deviance for the exponential relative to the Pareto.

The main shortfall of using minimum deviance as a tool for model selection is that it does not account for uncertainty about parameter estimates and differences in functional form complexity. The functional form of a model dictates the way in which parameters can interact. Different algebraic relationships between parameters in different models can lead to a differential ability of models with the same number of nominal parameters to fit noisy data patterns. What is needed is a way to penalize more complex models for the ability to fit random data patterns.

Hierarchical Bayesian Estimation

A hierarchical model adds the assumption that each participant, characterised by their parameter vector Θ_i for $i = 1 \ldots P$ participants, is a sample from a population distribution. In our application we assumed a multivariate normal population distribution, with parameters consisting of a vector of means, μ and a variance-covariance matrix Σ. The three means estimate the central tendency of the population. The Σ matrix consists of the three variances on the main diagonal, which estimates the extent of individual differences, and the three co-variances, which estimate the population correlations amongst parameters. We allowed for such correlations because it might be the case that, for example, participants with a generally better memory have both good initial encoding (b) and a slower rate of forgetting (α or β). In order to conform to the unbounded range of the normal, we estimated the probit transform of the a, $\acute{a}$ and b parameters and the logarithm of the positive parameters (α ,β and γ). The hierarchical models introduce another sort of functional form complexity related to the amount of shrinkage associated with a particular forgetting function. Greater shrinkage results in a less complex, and hence less flexible, model. This second type of functional form complexity must also be accounted for in model selection.

Although hierarchical models can be estimated by maximum likelihood (see Farrell & Ludwig, 2008) determining the likelihood of each data point requires an integration that can be difficult to perform in practice. Bayesian estimation using Markov Chain Monte Carlo (MCMC) methods provides an easy-to-implement alternative given the availability of general MCMC packages such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), which we used here. Informally, MCMC methods can be thought of as producing a set of population parameter samples, corresponding participant parameter and posterior predictive data samples (see Andrieu, DeFreitas, Douchet, & Jordan, 2003 for a comprehensive history and overview of MCMC methods).

Bayesian estimation requires a further set of assumptions, about prior distributions, which specify knowledge of model parameter values before the data are observed. For example, if nothing is known about a parameter except that it is on the unit interval, assuming a uniform prior is reasonable. For the results we report in detail later, we assumed a uniform prior for the population means of our a (for the exponential and power) and b parameters, which corresponds to a standard normal prior on the probit scale. For the population means of the remaining parameters (i.e., the logarithms of α, β, and γ) we assumed a normal prior with a mean of zero and standard deviation of 5. This prior is diffuse, in the sense of having appreciable mass over a broad range of parameter values, and has a median of one on the original scale for these parameters, which is close to typical estimated values. Finally, we made the convenient assumption of an inverse Wishart prior, $W^{-1}(m, \psi)$ for Σ (Tanner, 1998). For our 3x3 Σ matrix the inverse Wishart prior has parameters $m > 2$ and ψ, positive definite inverse scale matrix. We used the least informative value of m=3 and set ψ to the identity matrix. Figure 1 summarizes the Bayesian hierarchical model graphically (see Lee, 2008, for an introduction and examples of this notation) for the case of the exponential model. Note that hierarchical modelling does not require specification of covariance hyper-parameters. Potential correlations between parameters can be investigated by examining correlations between posterior parameter in a model assuming independence. When we did this we found sufficient correlation to warrant including explicit covariance parameters in our models. This has the advantage of providing improved estimates of parameter correlations as well as improving MCMC sampling efficiency. Essentially the same approach is used, for the same reasons, by Morey (in press) in studying different aspects of human memory.

The aim of MCMC estimation is to produce a sequence of samples from the joint posterior distribution of the parameters [5], where the posterior density of a parameter vector is proportional to its prior density times its likelihood given the data. Measures of

the central tendency of the posterior samples, such as the mean, provide an estimate of the population parameters. Variation among the samples reflects uncertainty about each parameter's true value. Hence, the quantiles of the posterior distribution can be used to construct parameter interval estimates, which are called "credible intervals" in Bayesian estimation. For example, the 2.5 % and 97.5 % quantiles define the end points of the 95 % credible interval.

We estimated the Bayesian hierarchical models described above separately for the explicit and implicit data sets from all lags. The lines in figures 2, 3 and 4 plot the posterior prediction of the model based on the expected posterior value of parameters in each of the models. Each panel in the figures also plots the same set of point and 95 % credible interval estimates of population retention probabilities. These estimates were calculated using Bayesian hierarchical models which did not assume a forgetting function; rather, they assumed a 10×10 multivariate normal population distribution (with an arbitrary variance-covariance matrix) of probit scaled retention. The multivariate normal transformation is the Bayesian equivalent plotting maximum likelihood estimates of group performance at each lag as is common in studies of retention. The technique yielded parameter vectors of population mean retention probability estimates. These estimates were averaged, and their 2.5 % and 97.5 % quantiles calculated, and both appropriately transformed to obtain the points and intervals in Figures 2, 3 and 4. Note that the ordinates in Figures 2, 3 and 4 have a $\log_{10}(lag)$ scale so that results for each lag can be easily distinguished.

The point estimates in figures 2, 3 and 4 indicate that retention in the explicit and implicit conditions differed only for shorter lags. The difference decreased with lag and was negligible after the first session. For both conditions retention decreased slightly from the end of session through to session three, but was essentially identical for lags of 7 and 28 days. Averell and Heathcote (2009) used interval estimates to argue that even at 28

days performance was well above chance. In their logit scaled analysis, and in the probit scale analysis reported here the proportion of samples falling below the chance completion rate was 0.0013 or less in all cases.

Figures 2, 3 and 4 show that Averell and Heathcote's (2009) design and results fulfil recommendations made by Rubin et al. (1999) for distinguishing amongst forgetting functions; that there be nine or more lags with a large ratio of longest to shortest lag, and that data points are away from ceiling and floor, with interval estimates that are precise, with a large ratio of most to least remembered. They recommended that functions that do not remain within the interval estimates be rejected. The power function comes closest to fulfilling this criterion, falling outside the 95 % credible intervals for both conditions only at the third lag. However, this method of model selection, even using the Bayesian intervals, does not take account of differences in model complexity. In the next section we apply model selection techniques that do make adjustments for complexity, although to varying degrees. We report results for several approaches following Liu and Aitkin (2008) suggestion that this provides another form of sensitivity check.

Bayesian Model Selection

Posterior deviance values for each MCMC sample $j = 1 \ldots M$ can be used as a basis for model selection (see Shiffrin et al., 2008 for discussion of alternative approaches). Each value is obtained by plugging each MCMC forgetting function parameter estimates for each participant, Θ_{ij} into their forgetting function and substituting the resulting retention probability estimates into the binomial deviance equation (5). These deviance values are summed over participants to produce the set of posterior deviance values, $D(\Theta_j)$

Two of our model selection methods (Raftery, Newton, Sagagopan, & Krivitski, 2007) AICM and BICM, require the deviance values to be independent. To achieve independence, we thinned our MCMC chains, retaining only one in every K values. Note

that the results reported previously were also based on these thinned chains. The value of K required, which varied between models, was indicated by examining autocorrelation functions and using the "effetiveSize" function provided by Plummer, Best, Cowles, and Vines's (2009) "coda" package for the R statistical language. The latter function determines the effective MCMC sample size adjusted for autocorrelation; we chose a value of K such that the thinned chain had an actual and effective size of 10,000. We found that this number of independent deviance values, was sufficient to reduce the BICM Monte-Carlo standard-error estimate provided by Raftery et al. (2007) to a level that did not introduce any ambiguity into the model selection results.

We examined three model selection "information criteria" calculated from Monte Carlo posterior deviance values. As well as the Monte Carlo Akakie (AICM) and Bayesian (BICM) information criteria mentioned previously, we also examined the more commonly used Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & Linde, 2002) Each of these criteria is based on the mean of the set of posterior deviance values, $\overline{D(\Theta_i)}$ and an estimate of the effective number of parameters in the hierarchical model. Differences in model complexity can cause estimates of the effective number of parameters to vary from the nominal number of parameters, which equals 48 for each of our three-parameter forgetting functions (i.e., 3×16, as 16 participant's data contributes to each hierarchical model).

For DIC, the estimate of the effective number of parameters is $p_D = \overline{D(\Theta_i} - D\overline{(\Theta_i)}$, where the latter term is a deviance calculated based on the average parameter values, $\overline{\Theta_i}$. The pD measure is sensitive to the constraint or shrinkage imposed by the hierarchical structure in the model (Gelman et al. 2004). If there is little constraint p_D divided by the number of participates will approximate the nominal number of forgetting function parameters. However, when there is constraint, the estimate of 'effective parameters' can differ from the nominal value. A major concern in hierarchical model selection is that the

hyper-distributions and their priors may impose different degrees of shrinkage for different models. Estimates of the hyper-distribution standard deviations and correlations can be used to examine the degree of shrinkage. It is also important to note that p_D $D(\overline{\Theta_i})$ and hence p_D is not parametrization invariant. In our application, for example, the value of $D(\overline{\Theta_i})$ differs depending on whether the average of Θ is taken on the probit and logarithmic scales used for estimation or on their original scales. The results we report here used the former scale however the model selection resultss do not differ if the later scale is used.

For the other criteria the estimate of the effective number of parameters is $p_V = \mathrm{Var}(D(\Theta_i))/2$. As the variance of the posterior deviance is parameterization invariant, so is p_V. More complex models have a posterior deviance distribution that is more variable. While the complexity penalty p_V is sensitive to the constraint imposed by the hierarchical structure Raftery et al. (2007) suggest that BICM is an asymptotic approximation of a Bayes factor so p_V is also sensitive to differences in the functional form complexity resulting from differences in the way parameters interact within a forgetting function. Note that for both estimates the effective number of parameters is not an absolute property of a model, it also depends on the data and the design from which they come (e.g., the lag values measured). Table 2 provides the estimates of the effective number of parameters per participant (i.e., $p_D/16$ and $p_V/16$) as well as the overall $\overline{D(\Theta_i)}$ values for each model in the explicit and implicit conditions based on all lags.

Both measures of the effective number of parameters indicate that the Pareto model is least complex and the exponential model most complex, with the power model intermediate. As variance is always positive, the p_V estimates are always positive, but this is not the case for the p_D estimate, which, as shown in Table 2, are negative for all models in the implicit condition. The negative p_D values for the implicit condition are problematic. Spiegelhalter et al. (2002) suggest that negative p_D values can be produced

from non-normal posterior distributions or when the model is not a good description of the data. We investigated these possibilities in our data and found neither were applicable. Further, estimates remained negative when using central tendency estimates (e.g., median or mode) other than the mean as well as averaging on different scales in the calculation of p_D and the same models were selected. Due to the negative p_D values we recommend caution when interpreting the DIC results for the implicit instruction condition.

Regardless ofthe negative pD values, model selection based on the three information criteria produced consistent results favoring the power function, as shown in Table 3. Each criterion adds to the mean posterior deviance a correction that is an increasing function of model complexity, so the model with the smallest value of the criterion is selected, DIC $= \overline{D(\Theta_i)} + p_D$, AICM $= \overline{D(\Theta_i)} + p_V$ and BICM $= \overline{D(\Theta_i)} + p_V \times \ln \sum_i N_i$. BICM applies a harshest complexity correction for all but very small data samples, and has been criticised for over correction (see Carlin and Spiegelhalter's discussion in Raftery et al., 2007 pp.33-36). The AICM and BICM results for a set of models can be transformed into weights making their values more interpretable as the conditional probability of each model (Wagenmakers & Farrell, 2004). These values are given in brackets in Table 3. In all cases the exponential model has negligible support. The AICM weights indicate very strong evidence in favor of the power model, whereas the BICM weights are more equivocal, but still clearly favor the power model.

By inspecting the hyper-distribution standard deviations we can gain an understanding of how much pooling is occurring across models. Larger standard deviation in the hyper-distributions equates to less constraint by the imposed hierarchical structure. Table 4 shows the hyper-distribution standard deviation for each model in both the explicit and implicit instruction conditions. The Pareto has a smaller standard deviation for the b parameter and overall lower standard deviation estimates in the explicit instruction condition, equating to lower p_D values. However, the lower complexity penalty

is not enough to make up for its misfit as reflected in its generally higher posterior deviance. The exponential and power models are roughly equivalent in the standard deviation of the hyper-parameter for the asymptote and scale parameters the rate parameter (α) standard deviation estimates in the rate parameter for the exponential is slightly larger than the rate parameter (β) estimates for the power. Therefore the lower DIC for the power model may be the result of differential shrinkage across models.

To further examine the possibility of differential shrinkage effecting the DIC results as well as to investigate the possibility of prior sensitivity in model selection (see Liu & Aitkin, 2008) we examined the effect of a range of priors; repeating our analyses with prior standard deviations of 2 and 1, respectively, for the probit and logarithmic scale population mean as well as a very diffuse set of priors where probit scale parameters were given a standard deviation of 2 while the logarithmic scaled parameters had a standard deviation of 5 [6]. We also analysed model selection with a range of values for ψ, the inverse Wishart hyper-prior. With all sets of hyper-priors the posterior variances did not change from the results in table 3 and again the power model was favoured by all model selection techniques in both experimental conditions. The outcomes suggest that the results reported in Tables 1 to 3 show little prior sensitivity. This includes the pD values for the implicit condition, which remained negative for all sets of hyper-priors.

Although the model selection techniques above all point to the power model supremacy at the hierarchical level it is also worth investigating model predictions against actual performance at an individual level. Posterior predictive distributions are useful in such a comparison and are generated based on equation 6.

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d(\theta) \tag{6}$$

Where y^{rep} can be thought of as values (in this case counts of correct completions at each lag) that would be observed if the conditions generating y were reintroduced. The integral

gives the probability density of y^{rep} given the values of θ as well as the posterior distribution of θ given the data y across parameter space $d(\theta)$ (see Lynch & Western, 2004 for further discussion). WinBUGS (Lunn et al., 2000) can compute posterior predictive distributions with the use of the cut function. We can compare these posterior predictive distributions to actual performance to asses model performance. Indication of a models inadequacies are seen where the posterior predictive distributions fail to capture trends in individual performance.

Figures 5 and 6 rrepresent the posterior predictive distributions at each lag as vertically aligned squares where the size of each of the squares represents the probability of each retention count. Observed performance is indicated by the black line (see Shiffrin et al., 2008 The results for participant 9 in the explicit condition and participant 15 in the implicit condition, shown in figures 5 and 6 respectively, are representative of results for other participants. In both figures it is evident that, relative to the power model, the exponential under-predicts performance at the later lags in session 1 (lags 6 and 7) and over-predicts performance at the later sessions (lags 8-10). The Pareto exhibits the opposite pattern, over-prediction at lags 6 and 7 in the first session and under-prediction for later sessions. The same trends are evident in the population level results illustrated in figures 2, 3 and 4.

General Discussion

The search for a general description of forgetting is one of the oldest unresolved problems in experimental psychology. We proposed that the difficulty in resolving this problem stems from issues relating to: 1) the level of measurement noise and the length of the retention period, 2) fitting models to data averaged over participants and 3) model selection techniques that do not account for differential complexity between candidate forms of the forgetting curve.

We addressed the first problem by analyzing data collected by Averell and Heathcote (2009), with a large number of observations per participant per retention interval, and retention measurements from one minute to 28 days. We avoided the second problem, while also minimizing measurement noise by analyzing data from all participants simultaneously, using hierarchal models estimated by Bayesian methods. Importantly the hierarchal models offer the level of psychological abstraction necessary to infer processes within the population without suffering the disadvantages distortion due to averaging. We addressed the third problem using Bayesian model selection techniques. These techniques required only information easily available from standard MCMC estimation, posterior deviance values. Consequently, both the estimation of hierarchical models and Bayesian model selection were accomplished relatively easily, making this approach readily available to other researchers.

Our analysis revealed that, although for individual participant data the exponential function with an above chance asymptote had the best fit among the models we considered, this advantage was due to its extra flexibility (complexity). When we adjusted for complexity using a range of model selection techniques that varied in the degree to which they adjusted for complexity, in every case a power function with an above chance asymptote provided the best description of forgetting. Interestingly, previous analyses of retention functions without an asymptote (Lee, 2004) found that the power model was more complex than the exponential. Our findings suggest that the addition of asymptote parameters adds more complexity to the exponential function than the power function.

The Power model of Forgetting

The power function was selected as the best forgetting curve for data collected under both explicit and implicit memory instructions. Table 5 shows the estimated estimated posterior parameter values and 95% credible interval for the power function.

The a and γ estimates are very similar, but the b parameter is slightly greater under for explicit than implicit, suggesting that instructions produced differences in initial performance, but that the rate at which participant's performance declined and the final level of retention were almost identical. The lower bound of the credible interval for the a parameter in both conditions is above the chance completion level of .116 indicating that the asymptote parameter was necessary. The correlation estimates in table 4 show mild departures from independence. The forgetting functions with an asymptote displayed a small positive correlation between the a and b parameters. This suggests that participants with a higher asymptote also have a greater estimated level of initial retention (i.e., $a + (1 - a) \times b$), perhaps due to individual differences in overall mnemonic ability. The correlations between the forgetting rate and asymptote was weak for both exponential and power functions, but there were larger positive correlations between b and the forgetting rate, and this was also true for the Pareto function. The latter correlations suggest that participants with a greater overall decrease in retention relative to their asymptotic performance forgot at a faster rate. Similarly larger, but negative, correlations occurred between the Pareto forgetting rate and γ parameters. This suggests that, particularly in the implicit condition, there was a trade-off between these parameters, whereas the Pareto b and γ parameters were largely independent.

The similarity in the predicted posterior parameter estimates for the explicit and implicit instruction conditions resemble those of McBride and Dosher (1997) (see also Dorfam, Kihlstrom, Cork, & Misiaszek, 1995), which they took to be suggestive of a single system underlying performance on both tasks. Kinder and Shanks (2001) provide a cognitive single system model of other phenomena used as evidence for separate explicit and implicit memory systems (but see Reber, 2002), and Wais, Wixted, Hopkins, and Squire (2006) suggest that the same hippocampual circuits underlie performance in both explicit and implicit memory tasks. Better initial performance under explicit instructions

may be due to a conscious effort to reinstate retrieval cues that are available within the first session but subsequently become unavailable. Consistent with this characterization, implicit and explicit performance was essentially identical in later sessions.

The ability of the power function to describe theoretical postulates believed common to forgetting, as well as a broad array of other cognitive processes, such as the relationship between perceptual magnitude and the judgment of that magnitude (Stevens, 1957), and the need to retrieve information in ecological settings (Anderson, 1990; Schooler, 1998) led G. D. A. Brown et al. (2007) to suggest that the power function be treated as a default model for cognitive processes until such time that sufficient evidence against it is found. The power law of forgetting has been used to describe forgetting at a neural level, where interference from other memory traces causes a breakdown in consolidation processes (Wixted, 2004a), but with a diminishing effect as retention increases, consistent with the power function's declining hazard rate (see Simmon, 1966). Although the findings presented above are consistent with such a consolidation processes, they are not in agreement with the Wixted's (2004a) conclusions regarding the ultimate fate of memories. Hence, if competition for consolidation is the cause of forgetting, it appears that ultimately some memory traces 'win' the competition and are permanently stored.

A power law of forgetting has also been attributed to a purely cue overload process (G. D. A. Brown et al., 2007). Specifically, a power model of forgetting can capture the buildup of interference where interfering material is assumed to be logarithmically compressed within cognitive space as a function of retention interval. Logarithmic compression of items in memory has the effect of making them increasingly confusable as time proceeds. The logarithmic compression of information is one of the assumptions of the SIMPLE (Scale Invariance Memory Perception and Learning model; G. D. A. Brown et al., 2007). However, the cue overload argument presented by G. D. A. Brown et al. (2007) also assumes a process that ultimately renders memories inaccessible due to the buildup of

interference, an assumption that is not in keeping with the results reported here. It should be noted that the two explanations need not be mutually exclusive, indeed Wixted (2004a) argued that the two causes of forgetting both occur in normal human functioning.

Within Session Effects

The power function was selected based data from retention periods extending across several experimental sessions over a 28 day period. However, many retention experiments are conducted within a single session. When we analyzed data from only the first session of Averell & Heathcote, 2009, results in favor of the power function were less convincing, and, overall, results were less consistent. Model selection results based on posterior likelihood ratios were equivocal in all cases. The exponential function was preferred by both AICM and DIC for the explicit data, and by DIC for the implicit data. The power function was preferred by AICM and BICM for the implicit data, and the Pareto function was preferred by BICM for the explicit data. Clearly, the latter result is questionable in light of all selection methods placing the Pareto function last with the full data set, as including longer lags should favor the Pareto function by measuring the very slowly declining performance which it can model. Hence this result is likely due to an over-correction for complexity by BICM (Spiegelhalter et al., 2002). On balance, however, the other findings indicate that the exponential function provides the best account of the session one data.

One possible account of these differences between first session model selection results and the results for all sessions is that, consistent with the multiple-scale nature of the power function, two processes with different time scales are acting to disrupt memory performance over the full 28 day period. The fast time scale process dominates forgetting within the first session, resulting in approximately exponential forgetting (also see Rubin et al., 1999, for evidence of an even faster time scale process that they identify with

short-term memory). Across later sessions the longer time scale process dominates, and so when data from all sessions are fit simultaneously the multi-scale power function provides a better account. Therefore, a power law of retention might not be an absolute, but instead depend on the length of the retention interval. However, as most attempts to retrieve information in the real world happen outside of the context in which the information was encoded, and often over time-frames of days, weeks and even years, a power function may represent the most ecologically valid quantitative description of forgetting.

Above chance asymptotes and Josts second law

The power law of forgetting quantifies one of the oldest verbal 'laws' in experimental psychology, Jost (1897) second law. Our results partially support Jost's second law. Our findings suggest that if two memory traces are equal in strength, but sufficiently different in age, the younger one will decline faster than the older one, due to the influence of the fast time-scale process on the younger but not older trace. However, this law only holds until both traces have reached asymptote. Obviously, at this stage, the age of the trace is irrelevant as both traces are now no longer declining. The presence of an above chance asymptote for the power model suggests that some memory traces are resistant to the force imposed by other memories either at the neural or cognitive level. Given that 28 days is a sufficient period to address the concerns expressed by McBride & Dosher, 1997, as well as other similar concerns (e.g. Rubin et al., 1999) about declines too small to detect within a single session, our results agree with Chechile's (2006) suggestion that not allowing for the possibility of permanent retention constitutes a "serious failing" (p.36). Results in favor of an above chance asymptote are particularly bolstered by consistent selection of the Pareto of the worst model of forgetting over 28 days, given the parametric flexibility of this function to model very slow declines.

There are both cognitive and neural mechanisms that could support the permanent

storage of memory traces. From a neurological perspective Arshavsky (2006)suggests that memory traces that survive long enough become stored as structural changes in DNA and are therefore permanent. Interestingly, Arshavsky (2006) believes that the process by which changes in long term potentiation are transferred into changes in DNA happens over the first few weeks after encoding. Therefore, the above chance asymptotic performance seen here could be the result of structural changes in DNA. The DNA hypothesis is attractive because it offers a solution to the problem of capacity. Alternative neural hypothesis regarding memory formation, such as structural changes at the synapse of neurons, have a limited capacity in terms of the overall number of memories that can be stored. However, structural changes in DNA would allow for an almost limitless number of memories to be permanently stored.

From a cognitive perspective our results suggest that some memories remain free from the detrimental affects of interference (i.e., they stand out from the noise resulting from the logarithmic compression of memory traces), and it is this distinctiveness that is driving the above chance asymptotic performance seen in the results. Retrieval cues provided by the environment at test may provide a mechanism that reduces the interference. The experiment examined here offered both item cue support (the first three letters of the critical word) and environmental cue support (video tape and questionnaire). If forgetting is driven largely by interference with retrieval by previous and intervening material, than the retrieval cue support given in this experiment may have alleviated the effects of interference, thereby allowing more of the stored information to be translated into performance. Accordingly, the asymptote may correspond to the amount of retrieval support given and, as suggested by Rubin et al. (1999), the asymptote parameter may be useful in analysis of retention data until the experimental context at test is totally different from the experimental context at study.

Author Note

Address correspondence to Lee Averell: Aviation Building, University Avenue, Callaghan University of Newcastle Fax: (61-2) 49216906 NSW, 2308, AUSTRALIA Office: (61-2)49216959 lee.averell@newcastle.edu.au

References

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

Andrieu, C., DeFreitas, N., Douchet, A., & Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine Learning*, *50*, 5-43.

Arshavsky, Y. I. (2006). "the seven sins" of the hebbian synapse: Can the hypothesis of synaptic plasticity explain long-term memory consolidation? *Progress in Neurobiology*, *80*, 99-113.

Averell, L., & Heathcote, A. (2009). Long term implicit and explicit memory for briefly studied words. In A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual confrence of the cognitive science society* (p. 267-281). Austin TX: Cognitive Science society.

Bahrick, P., H. (1987). Semantic memory content in permastore: Fifty years of memory for spanish learned in school. *Journal of experimental psychology; General*, *113*, 1-26.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539-576.

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

Chechile, R. A. (2006). Memory hazard functions: A vechile for theory development and test. *Psychological Review*, *113*, 31-56.

Cohen, A., Sandborn, A., & Shiffrin, R. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin and Review*, *15*, 692-712.

Dorfam, J., Kihlstrom, J. F., Cork, R. C., & Misiaszek. (1995). Priming and recognition in ect induced amnisia. *Psychonomic Bulletin and Review*, *2*, 224-248.

Ebbinghaus, H. (1885/1974). *Memory: A contribution to experimental psychology.* New

York: Dover.

Farrell, S., & Ludwig, C. H. (2008). Bayesian and maximum likelihooh estimation of hierarchical response time models. *Psychonomic Bulletin and Review*, *15*, 1209-1217.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269-299.

Jost, A. (1897). Die assoziationsfestgkeit in iher abhangigkeit von der verteilung der wiederholungen [the strength of association in their dependence on the distribution of representations]. *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane*, *16*, 436-472.

Kinder, A., & Shanks, D. R. (2001). Amnesia and the declarative/non-declarative distinction: A recurrent network model of classification, recognition and repetition priming. *Journal of cognitive neuroscience*, *13*, 95-105.

Lee, M. D. (2004). A bayesian analysis of retention function. *Lournal of mathematical psychology*, *48*, 310-321.

Lee, M. D. (2008). Three case studies in the bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, *15*, 1-15.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalisabilty. *Journal of Mathematical psychology*, *52*, 362-375.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs a bayesian modelling framework: concepts, structure and extensibility. *statistics and Computing*, *10*, 325-337.

Lynch, S., & Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods and Resrarch*, *32*, 301-335.

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

Morey, R. (in press). A baysesian hierarchical model for the meausurement of working memory capacity. *Journal of Mathematical Psychology*.

Myung, I. J., & Pitt, M. A. (1997). Applying occam's razor in modeling cognition: A bayesian approch. *Psychonomic Bulletin and Review*, *4*, 79-95.

Myung, I. J., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychogical Review*, *116*, 499-518.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assesing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47-84.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2009). *Output analysis and diagnostic for mcmc* (Tech. Rep.). CRAN.

Raftery, A. E., Newton, M. A., Sagagopan, J. M., & Krivitski, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, *8*, 1-45.

Reber, P. J. (2002). Attempting to model dissociations in memort. *Trends in cognitive neuroscience*, *6*, 192-194.

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual review of psychology*, *59*, 225-254.

Rubin, D. C., Hinton, S., & Wenzel, A. E. (1999). The precise time course of forgetting. *Journal of experimental psychology*, *25*, 1161-1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *203*, 734-760.

Schooler, L. (1998). Sorting out core memory processes. In N. Chater & M. Oaksford (Eds.), *Rational model of cognition* (p. 128-155). Oxford: Oxford University Press.

Shiffrin, R. M., Lee, M. D., Wagenmakers, E.-J., & Kim, W. J. (2008). A survey of model evaluation approches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*, 1248-1284.

Simmon, H. A. (1966). A note on jost's law and exponential forgetting. *Psychometrika*, *31*, 505-506.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical society*, *64*, 583-639.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153-181.

Tanner, M. A. (1998). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions.* New York: Springer.

Wagenmakers, E. J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic Bulletin and Review*, *1*, 192-196.

Wais, P. E., Wixted, J. T., Hopkins, R., & Squire, L. R. (2006). The hippocampus supports both the recollection and the familiarity componants of recognition memory. *Neuron*, *49*, 459-466.

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

Wixted, J. T. (2004a). On common ground: Jost's(1897) law of forgetting and ribot's (1881) law of retrograde amnesia. *Psychological Review*, *111*, 864-879.

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

Footnotes

[1]Throughout this paper we will use the term 'fit' to mean a traditional goodness-of-fit measure of a particular set of parameters to data rather than in a Bayesian sense of where 'fit' is used to describe model adequacy of all possible combinations of model parameters

[2]Other causes might also apply, such as study resulting in encoding of a short-term memory representation but not a long-term memory representation. In this case measured retention might be perfect immediately after study, due to retrieval from short-term memory, even when $b < 1$. In Averell and Heathcote's (2009) experiment the interval between study and the first lag was filled with other study and test events, so retrieval from short-term memory was unlikely.

[3]We do not consider either the linear or logarithmic functions examined by Lee (2004) as they can make predictions outside the unit interval, and so are not suitable for retention probability data.

[4]An alternate approach to these questions involves examining a four-parameter Pareto function with an estimated asymptote. However, analyses with this function tended to be numerically unstable, often producing extremely small estimates of γ and correlated very large estimates of b. The reason is related to the Pareto's hazard function, which can be close to constant, like that of the exponential, over the range of experimentally measured lags when γ is small. Correlated large values of b compensate for the attendant very small change in $P(t)$ over the measured range of t.

[5]The raw sequence of samples or "chain" produced by MCMC takes some time to converge to the posterior distribution, and is often auto-correlated, which can cause a variety of problems. Typically initial samples before convergence are discarded, but very strong autocorrelation can cause the sequence to fail to converge. We report results based on single chains, which, although strongly auto-correlated, did converge after we discarded the first 20,000 samples. This was confirmed by visual inspection of the chain and checks

using multiple chains tested with GelmanRubin1992's statistic.

[6]We attempted to obtain MCMC samples with even more diffuse hyper-priors but WinBUGS frequently crashed at these levels

Table 1

Mean posterior deviance for each retention curve model,$D(\bar{\Theta}_i)$, and estimates of the effective number of parameters per participant, p_D and p_V, for implicit and explicit conditions based on data from all lags.

	Explicit			Implicit		
Model	$\overline{D(\Theta_i)}$	p_D	p_V	$\overline{D(\Theta_i)}$	p_D	p_V
Exponential	1011	2.43	3.04	923	-0.8	2.45
Power	937	1.66	2.74	879	-0.96	2.25
Pareto	1002	1.04	2.31	933	-2.42	1.87

Table 2

Mean posterior deviance for each retention curve model,$D(\bar{\Theta}_i)$, and estimates of the effective number of parameters per participant, p_D and p_V, for implicit and explicit conditions based on data from all lags.

	Explicit			Implicit		
Model	$\overline{D(\Theta_i)}$	p_D	p_V	$\overline{D(\Theta_i)}$	p_D	p_V
Exponential	1011	2.43	3.04	923	-0.8	2.45
Power	937	1.66	2.74	879	-0.96	2.25
Pareto	1002	1.04	2.31	933	-2.42	1.87

Table 3

Information criteria for implicit and explicit conditions based on data from all lags. conditional model probabilities based on AICM and BICM are given in brackets.

	Explicit			Implicit		
Model	DIC	AICM	BICM	DIC	AICM	BICM
Exponential	1049	1059(0)	1427(0)	910	962(0)	1260(0)
Power	963	981(1)	1313(.95)	864	915(1)	1188(.67)
Pareto	1019	1039(0)	1319(.05)	895	963(0)	1190(.33)

Table 4

Mean estimates of the hyper-distribution standard deviation (SD) and correlation (r) parameters for the exponential, power and Pareto models forexplicit and implicit instruction conditions.

	Explicit				Implicit		
Exponential	a	b	α	Exponential	a	b	α
SD	.32	.54	.7		.32	.55	.73
r	a,b	a,α	b,α		a,b	a,α	b,α
	.23	.04	.41		.24	-.09	.32
Power	a	b	β	Power	a	b	β
SD	.31	.6	.45		.31	.66	.53
r	a,b	a,α	b,β		a,b	a,β	b,β
	.14	.07	.18		.21	.052	.27
Pareto	b	γ	β	Pareto	b	γ	β
SD	.39	.4	.13		.7	1	.38
r	b,γ	b,β	γ,β		b,γ	b,β	γ,β
	.18	.3	-.37		.07	.27	-.7

Table 5

Mean estimated posterior parameter values (with 95% credible interval) for the power model in both explicit and implicit conditions

	Parameter		
	a	b	β
Explicit	.19 (.15,.24)	.78 (.71,.85)	.68 (.5,.9)
Implicit	.19 (.14,.24)	.62 (.43,.86)	.67 (.43,1.1)

Figure Captions

Figure 1. Graphical notated hierarchical model for the asymptote exponential model of forgetting. Graphical models can show the relationship between and among observed and unobserved variables in a model in such a way that allows for quick and easy viewing of the total model structure. In graphical models, nodes are used to represent variables and dependencies are built into the graph configuration itself (where second order or 'child' nodes depend on first order or 'parent' nodes). Here we use accepted convention, representing continuous variables with circular nodes and discrete variables as square nodes. Further, observed variables are shaded and unobserved variables unshaded. Stochastic variables are denoted with single boarders and deterministic variables have double boarders. In this model idividual participant parameter vectors Θ are drawn from a multivariate normal hyper-prior with mean μ and a $k \times k$ variance covariance matrix Σ which was assumed to have an inverse Wishart W^{-1} hyper-prior distribution. In this moel Δ represents the combination of μ and Σ for each node in the model. The unit interval parameters a_i and b_i correspond to probit transformed standard normal distributions. The logerithmic parameter α has a mean of zero and a standard deviation of 5. The hierarchical model model has the advantage over non-hierarchical having participants estimates modeled from a higher more abstract level. B=Binomial, MV N=multivariate normal, t= lags 1-10, i=participants.

Figure 2. Exponential model fits to both the explicit and implicit data. The points represent the population mean retention probability estimates. The error bars represent the 95% credible intervals for the population. Ticks on the ordinate indicate lags on a log10 scale. The vertical dot dash line at the bottom of the plot represents chance completion probability.

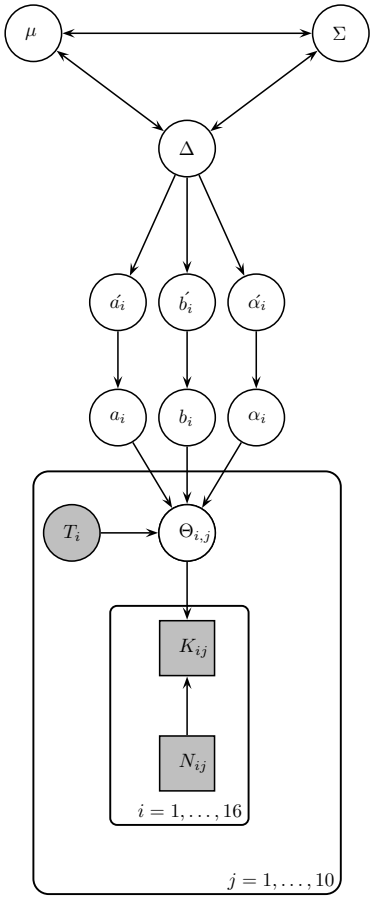
Figure 3. Power model fits to both the explicit and implicit data. The points represent

the population mean retention probability estimates. The error bars represent the 95% credible intervals for the population. Ticks on the ordinate indicate lags on a log10 scale. The vertical dot dash line at the bottom of the plot represents chance completion probability.

Figure 4. Pareto model fits to both the explicit and implicit data. The points represent the population mean retention probability estimates. The error bars represent the 95% credible intervals for the population. Ticks on the ordinate indicate lags on a log 10 scale. The verticlal dot dash line at the bottom of the plot represents chance completion probability.

Figure 5. Posterior predictive distribution for the power (panel 1), exponential (panel 2) and Pareto (panel 3) for participant 9 in the explicit condition. The vertically aligned squares represent the posterior mass of stems completed at each lag given the models assumptions. The black line represents counts of stems correctly completed at each lag for participant 9.

Figure 6. Posterior predictive distribution for the power (panel 1), exponential (panel 2) and Pareto (panel 3) for participant 15 in the implicit condition.



$K_{i,j} \sim \text{Binomial}(\theta_j, N_{i,j})$

$\mu \sim \text{MVNmean}(\mu, \sigma)$

$\Sigma \sim \text{W}^{-1}(m, \psi)$

$\Delta \sim \text{MVN}(\mu, \Sigma)$

$a_i' \sim \phi\text{Normal}(\mu, \Sigma)$

$b_i' \sim \phi\text{Normal}(\mu, \Sigma)$

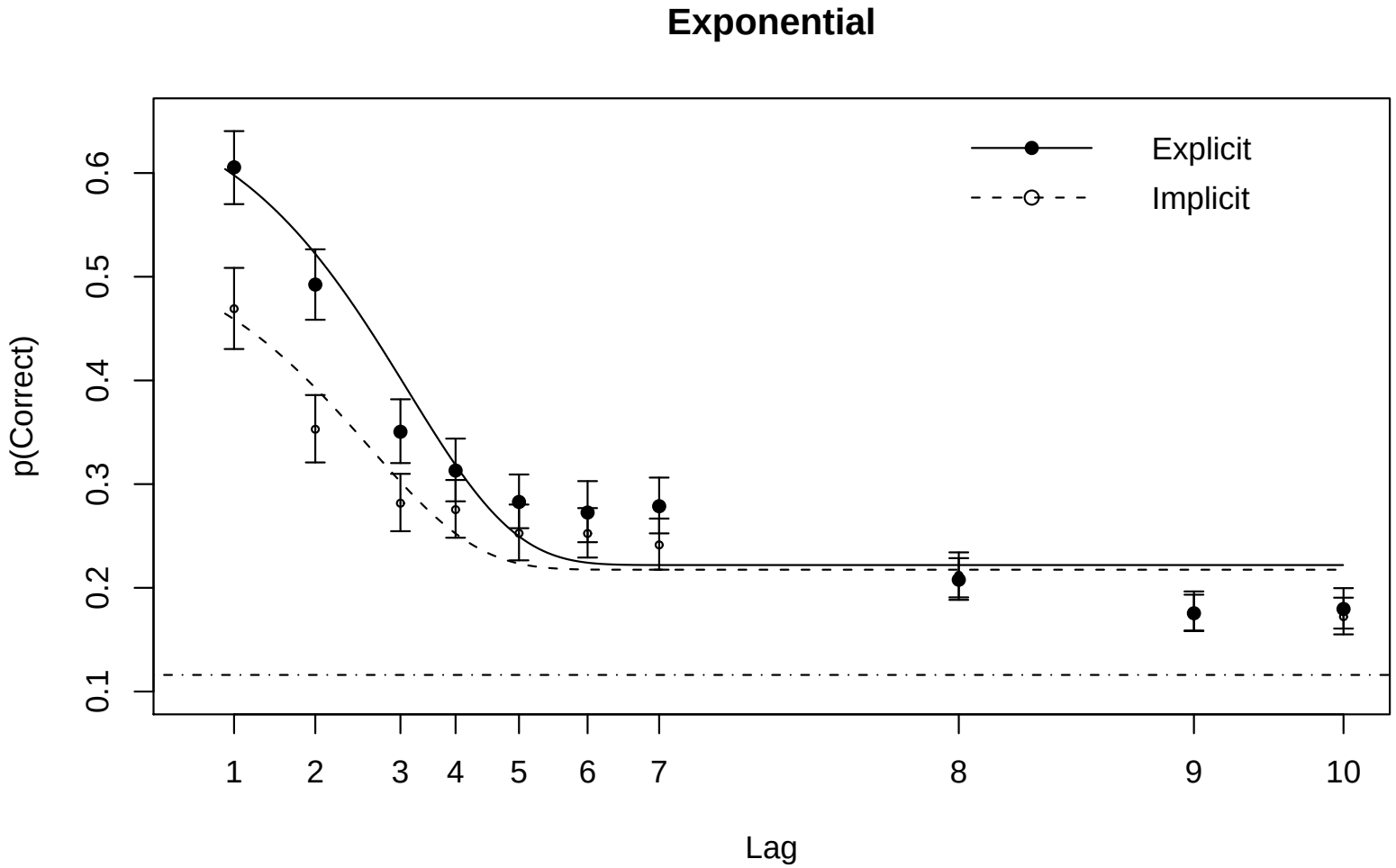
$\alpha_i' \sim \text{Normal}(\mu, \Sigma)$

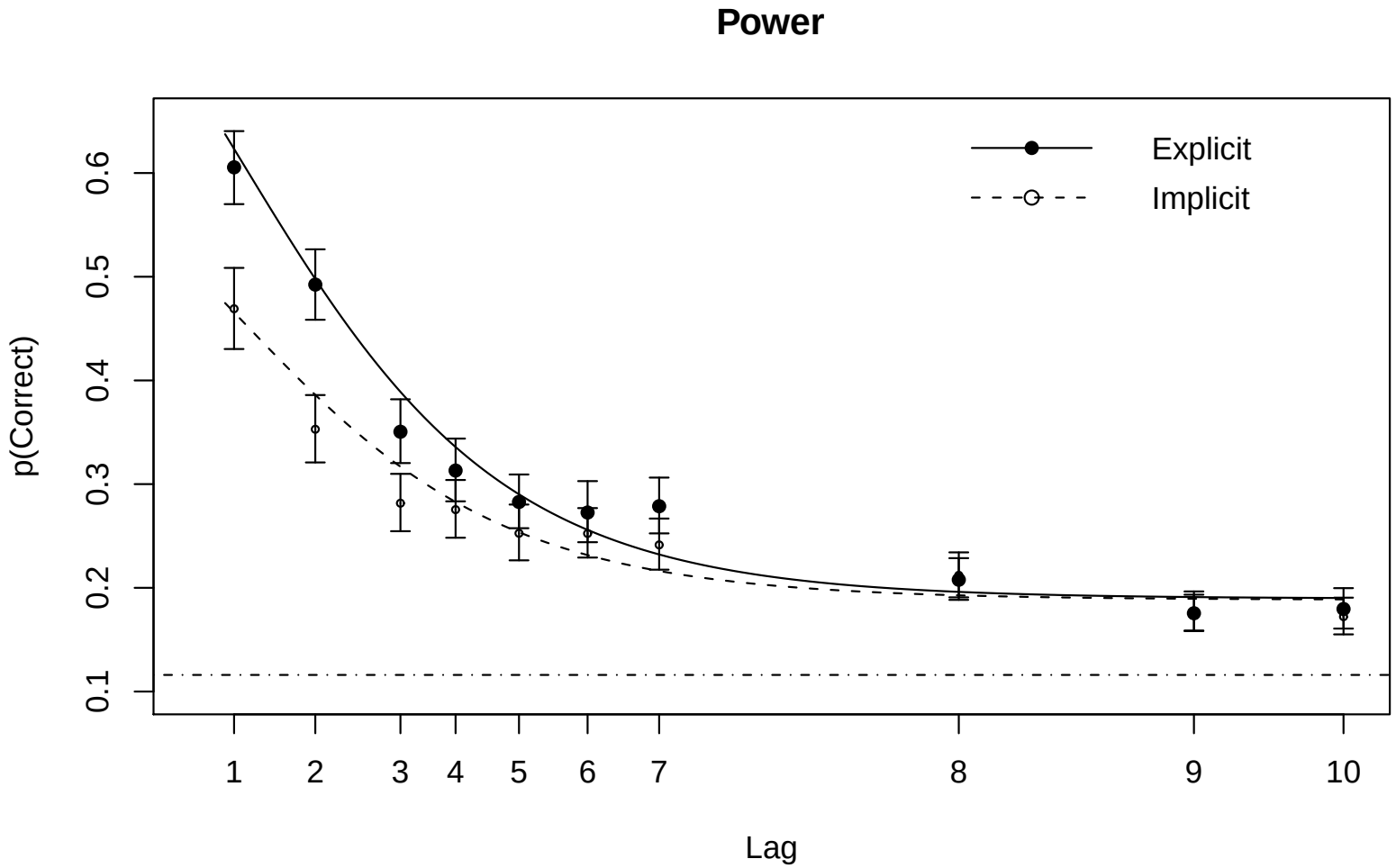
$a_i \sim \text{Uniform}(0, 1)$

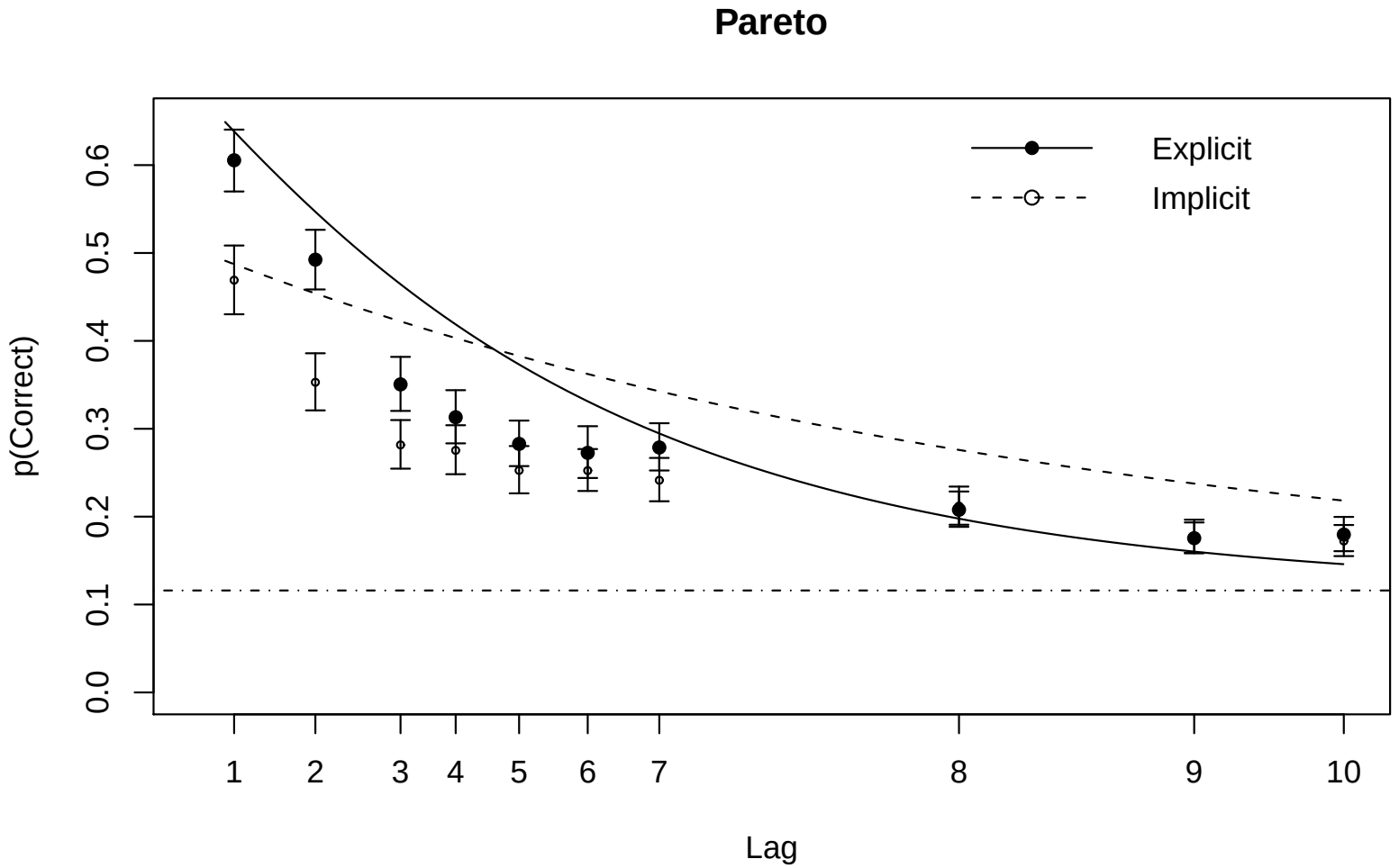
$b_i \sim \text{Unifom}(0, 1)$

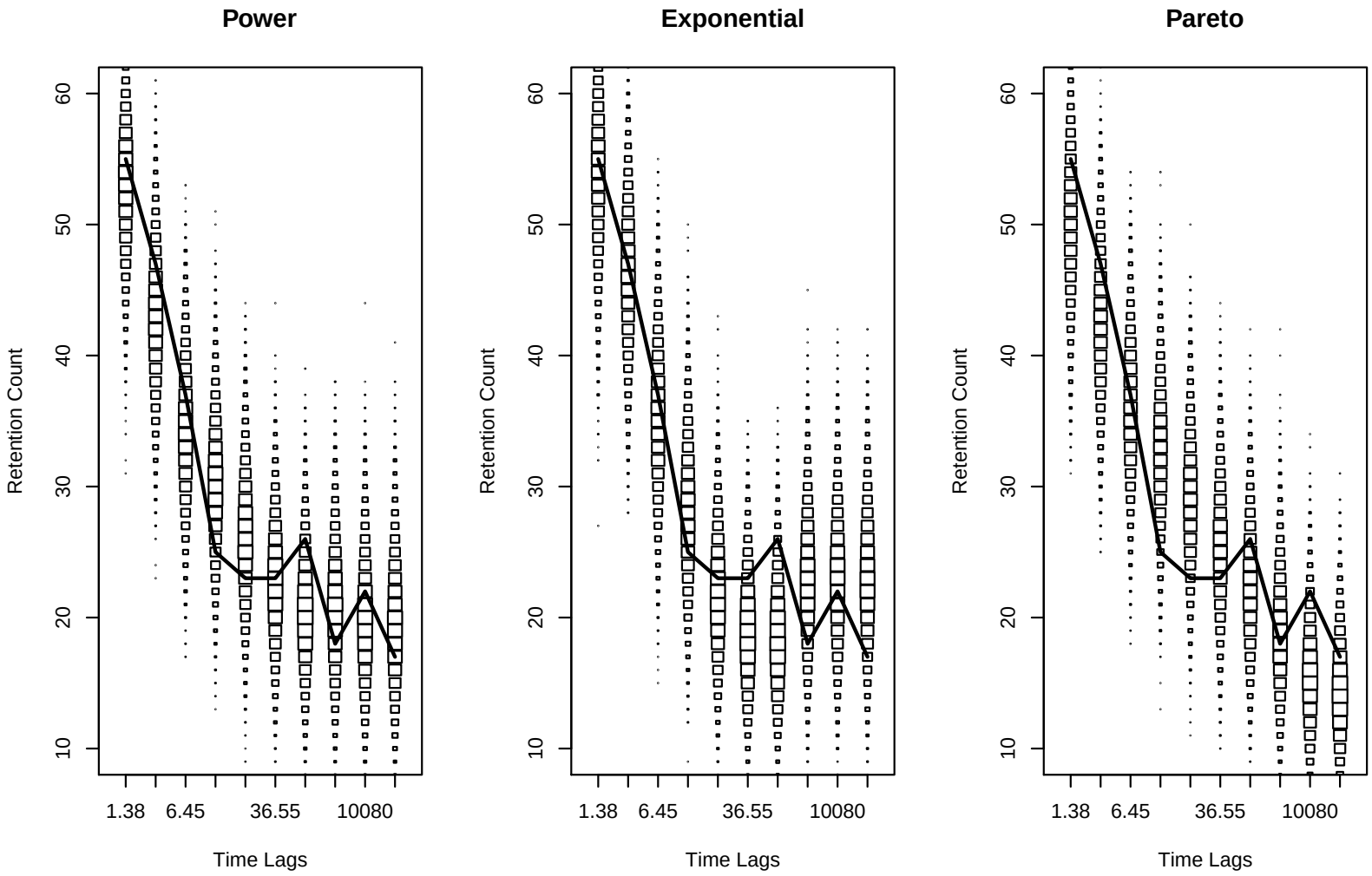
$\alpha_i \sim \text{LogNormal}(0, 5)$

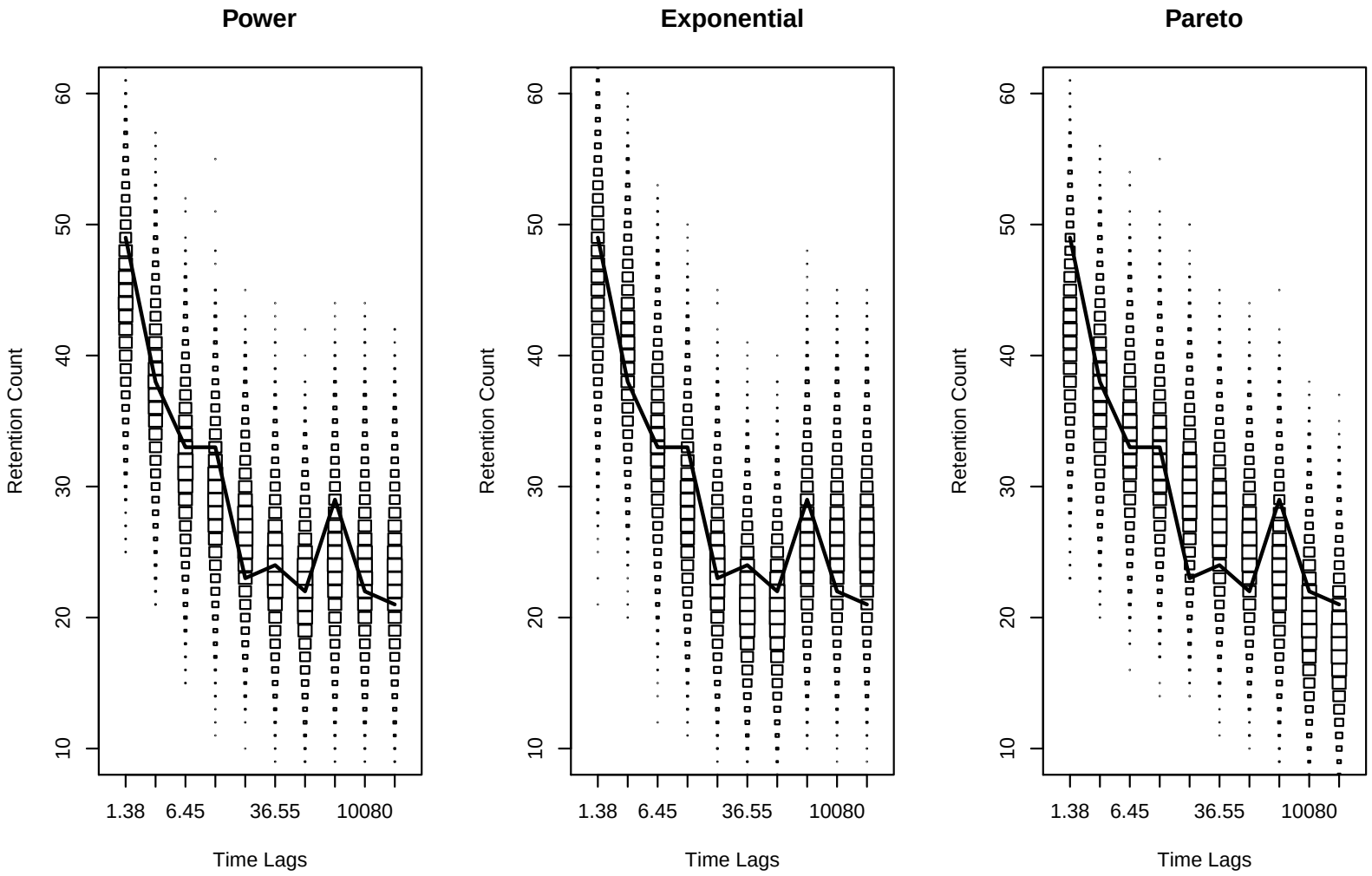
$\theta_{i,j} = .116 + (1 - .116) \times (a + (1 - a) \times b \times \exp^{(-\alpha t)})$


Exponential
Explicit
Implicit
p(Correct)
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
Lag


Power
Explicit
Implicit
p(Correct)
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
Lag


Pareto
Explicit
Implicit
p(Correct)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
Lag


Power
Exponential
Pareto
Retention Count
10
20
30
40
50
60
1.38
6.45
36.55
10080
Time Lags


Power
Exponential
Pareto
Retention Count
10
20
30
40
50
60
1.38
6.45
36.55
10080
Time Lags