A Bayesian Race Model for Recognition Memory

Sungmin Kim[1], **Kevin Potter**[2],

Peter F. Craigmile[1], **Mario Peruggia**[1], **Trisha Van Zandt**[2,1,*].

[1] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

[2] Department of Psychology, The Ohio State University, Columbus, OH 43210, USA

[*] Email: van-zandt.2@osu.edu

Abstract

Many psychological models use the idea of a trace, which represents a change in a person's cognitive state that arises as a result of processing a given stimulus. These models assume that a trace is always laid down when a stimulus is processed. In addition, some of these models explain how response times (RTs) and response accuracies arise from a process in which the different traces race against each other.

In this article we present a Bayesian hierarchical model of RT and accuracy in a difficult recognition memory experiment. The model includes a stochastic component that probabilistically determines whether a trace is laid down. The RTs and accuracies are modeled using a minimum gamma race model, with extra model components that allow for the effects of stimulus, sequential dependencies, and trend. Subject-specific effects, as well as ancillary effects due to processes such as perceptual encoding and guessing, are also captured in the hierarchy. Predictive checks show that our model fits the data well. Marginal likelihood evaluations show better predictive performance of our model compared to an approximate Weibull model.

Keywords: Cognitive modeling; Human performance data; Minimum gamma; Mixture modeling; Weibull.

1 Introduction

An old and important concept in models of human cognition is that of the "trace," dating as far back as Aristotle (Bloch, 2007). A trace is a change in the state of the cognitive system resulting from the presentation of a stimulus and possibly the response to it. The trace concept is central to many important models of cognition. Influential memory models use the idea of the trace to explain effects of multiple presentations of stimuli on recognition and recall performance (e.g., Shiffrin and Steyvers, 1997). Exemplar models of signal detection and categorization assume that novel stimuli are stored as points in a mental representation of the physical stimulus space, a non-Euclidean space whose axes are determined by the characteristics of the stimulus (e.g., shape, color, size) most important for identification (e.g., Nosofsky, 1987). Theories of learning and skill acquisition assume that each time a task is practiced a new trace is laid down (or an old trace is strengthened or modified), and as more traces are laid down the person becomes more expert in performing the task (e.g., Logan, 1988; Taatgen et al., 2008).

Another concept that is central to many important models of cognition is that of the race. A race represents a response to a stimulus as the result of a competition between all possible responses that could have been made. Information favorable toward each possible response accumulates over time until one response attains an amount sufficient to be selected. Different models of choice characterize the process of information accumulation differently (Ratcliff and Smith, 2004). Some models assume the accumulation mechanisms for different responses are correlated (Ratcliff, 1978; Usher and McClelland, 2001), and some assume they are independent (Van Zandt et al., 2000; Logan et al., 2014; Rouder et al., 2014). Some models are based on the general idea that information accumulates over time without specifying the nature of that information, while others make assumptions about where that information might come from and how it might change in different experimental conditions. In particular, several models propose that traces established during exposures to stimuli race against each other to be retrieved from memory, and the outcomes of this race are either agglomerated into a single information quantity that is then used as evidence toward a response (Logan, 1992; Nosofsky and Palmeri, 1997), or that these outcomes themselves determine the response (Logan, 1988).

Whether responses are based on a race between traces or on accumulation of information derived from the retrieval of such traces, the race processes describing retrieval and information accumulation take time. These models can therefore predict not only the final response that is made, but also the response times (RTs): the time between the onset of a stimulus and the response made to it. Human performance data, therefore, consist of both the set of RTs and the responses (accurate or not) made to the stimuli presented in an experiment.

In this paper, we focus on models of choice that assume the response is determined at least in part by a race between traces, a structure that characterizes a good many exemplar and skill-acquisition models. An interesting characteristic of most, if not all, trace-race models is that they assume that the process by which traces are laid down is deterministic. That is, with probability one, a trace is laid down for each stimulus presentation (Hintzman, 1988; Logan, 1988; Raaijmakers and Shiffrin, 1981). Some models assume that traces can decay with time (Nosofsky et al., 1992; Turner et al., 2011), but most memory and skill acquisition models assume that a (potentially corrupted) trace exists for each stimulus presentation and always contributes to each response through the course of an experiment. The appropriateness of this assumption depends on whether or not an experimental subject is actually paying attention during the task, and on whether or not the cognitive system can reliably lay down a trace each time a stimulus is presented.

Our article deals with the question of identifying the number of traces laid down during the performance of a recognition memory task. Earlier work by Craigmile et al. (2010b) developed a modeling framework within which we can estimate the probability that a trace is established. The model we present here is not a serious contender to more theoretically-nuanced and established models of memory, categorization or skill acquisition. Instead, we use a simplified model to characterize the trace-race process, a model that we can use to examine more carefully the trace concept and how traces are constructed.

Using the insight so gained into the trace construction process, we can identify subjects who do not build new traces during the performance of a task. That is, we can determine those subjects who were performing the task (performing subjects) and those who were not (underperforming subjects). Pre-processing of human performance data frequently requires researchers to examine a summary statistic describing overall performance (e.g., response accuracy) and to discard subjects on the basis of whether or not they meet some usually arbitrary criterion (e.g., greater than 65% accuracy). Without discarding any data our procedures allow performing and underperforming subjects' data to be modeled simultaneously. This not only prevents experimenter bias from influencing the results and conclusions, but also provides more insight into the cognitive process by telling us how it fails for underperforming subjects and succeeds for performing subjects.

In Section 2 we present data from a recognition memory experiment. In Section 3 we perform an exploratory analysis of the data. This includes a discrimination between performing and underperforming subjects using a popular statistic called d', and the assessment of the effect of the experimental manipulations on RT and accuracy. In Section 4 we introduce a hierarchical Bayesian model that incorporates a stochastic trace mechanism. We fit the model and summarize the posterior distributions of its parameters in Section 5. In Section 6 we discuss model validation and compare the predictive performance of our model to an approximate Weibull model. Our discussion appears in Section 7. Further details about the experiment, the data, the priors, and the Markov chain Monte Carlo algorithm used to fit the Bayesian models are provided in supplemental online material.

2 The Experiment

We decided that a recognition memory experiment would provide data most useful for developing a model and exploring characteristics of the trace process. A recognition memory task proceeds in two phases: In the initial study phase, a list of stimulus objects (words, pictures, etc.) are exposed a number of times. In the subsequent test phase, a subset of the objects presented during the study phase is selected, together with a set of new objects that were not presented, and together the old and new objects form a test list. In the test phase, subjects respond “old” or “new” to each test object, depending on whether or not they recognize it from the study phase.

Following other similar work (e.g., Musen and Treisman, 1990; Province and Rouder, 2012; Voss et al., 2010), we designed a recognition memory experiment in which abstract pictures were presented one or more times during the study phase. It should be noted that the memory task we designed for this experiment is unusually difficult. We had three reasons for designing the task in this way. First, we wanted a task in which the construction of a memory trace with repeated presentations was potentially random: the probability of constructing a new trace could be less than one. Second, we wanted to use stimuli that were completely unfamiliar and devoid of semantic content, reducing the opportunity for subjects to assign meaning to them, to be able to associate them with each other, or to be able to retrieve some of them from memory more easily than others. The unfamiliarity of the stimuli also meant that there would be no latent traces that could compete with the traces established during the experiment (i.e., there is minimal "context noise"; Dennis and Humphreys, 2001). Third, the difficulty of the task meant that some subjects could not do it at all, which provided an important modeling challenge.

The supplemental material gives a detailed description of our methods. Briefly, 32 young adults (12 in a pilot study and 20 in the final study) saw pictures of geometric objects arranged in novel configurations (see Figure 1) presented one at a time. During study, each of 8 pictures was exposed from 1 to 4 times. In the test phase, the subjects attempted to discriminate between the 8 old pictures that were studied and a randomly-selected group of 8 new pictures that were not studied. Over 21 study/test blocks, each subject produced a series of $16 \times 21 = 336$ responses, 168 responses to old pictures and 168 responses to new pictures. For each response, we measured the time to make the response (RT) and the response itself ("old" or "new"). We excluded the first block from the analysis to ensure that the subjects had become familiar with the task. We also excluded responses to the first two and the last two items on the study list (which were exposed only once) to avoid primacy and recency effects (Ebbinghaus, 1913; Murdock Jr, 1962). The removal of these buffer trials left a total of 240 responses: 80 responses to old pictures (20 for each number of exposures between 1 and 4) and 160 responses to new pictures for each subject.

3 Exploratory data analysis

We first performed an exploratory analysis to investigate the main features of the data and to ensure that any modeling assumptions we made would be sufficient to capture those features. In particular, we wanted to verify that the effect of number of exposures on RTs and response accuracy was consistent with similar effects in other recognition memory experiments.

3.1 Discriminability and d'

The sensitivity index d' is a measure of discrimination ability used in signal detection theory (Green and Swets, 1966) that has wide application in memory research. The use of d' assumes the existence of a single dimension that can be used to discriminate between stimulus types (old vs. new). In recognition memory, this dimension represents familiarity. If an object presented during the test phase has sufficiently high familiarity then it will be identified as old. The simplest signal detection model assumes that, across all presented objects, the distributions of familiarity for old and new objects are normal with a common variance (usually 1). Without loss of generality, we set the mean of the distribution of familiarity for new items equal to 0, and the mean for old items is then estimated as $d' = \Phi^{-1}(\mathrm{H}) - \Phi^{-1}(\mathrm{FA})$,where $\Phi^{-1}(\cdot)$ denotes the inverse of a standard normal cumulative distribution function, H is the “hit” rate, the proportion of studied pictures correctly identified as old, and FA is the “false alarm” rate, the proportion of unstudied pictures incorrectly identified as old. Higher values of d' indicate a greater separation between the mean perceived familiarities of old and new pictures, and hence a better ability to discriminate between them.

The d' statistic is an important measure of recognition memory performance (Egan, 1958; Banks, 1970; Yonelinas, 1994). In particular, as the number of exposures of a picture increases, its familiarity should increase, resulting in larger values of d' (e.g., Hirshman, 1995). In addition, the d' statistic can be used to discriminate between subjects who perform a task and those who do not. Data from subjects with low values of d' are often discarded on the assumption that they failed to follow instructions (e.g., Kates et al., 2007).

We divided the subjects into two groups (performing and underperforming) based on whether their overall d' (collapsed over number of exposures) was greater or less than 0.3. Seven out of the 20 subjects were classified as underperforming according to this criterion, which was motivated by the variance of the sampling distribution of d': if the old and new familiarity distributions have equal means (are indiscriminable), the central 90% of d' extends approximately from -0.3 to 0.3 for the sample sizes that we used here.

Next, we computed d' as a function of exposures, contrasting the hit rates for 1, 2, 3 and 4 exposures to the false alarm rate. Figure 2 plots d' against the number of exposures E separately for performing and underperforming subjects, and demonstrates two points. First, in general, performing subjects have d's that increase more rapidly with the number of exposures E. This increase is consistent with past memory research and demonstrates that varying the number of exposures had the desired effect. Second, using d' to identify performing and underperforming subjects may not accurately categorize subjects' performance. Some underperforming subjects (for example, Subjects 6 and 20) have d's that increase with E, but not as rapidly as for performing subjects. This may indicate that these subjects may simply have had more difficulty with the task than others with higher overall d'.

Removing 7 of 20 subjects' data for failure to follow instructions would mean discarding a third of the data collected in the experiment. It is not surprising that so many subjects performed poorly: the task itself is very difficult, unusually so for a memory experiment, and the subjects, being student volunteers, had little motivation beyond good will to extend their best efforts. Although it is standard practice to remove such data, we argue that a better strategy is to fit models to all the data, models that are flexible enough to identify underperforming subjects, and in this way isolate "good" from "bad" data in a principled and objective way. This was one of our main modeling challenges.

3.2 RTs as time series

RT data are nonstationary and exhibit sequential dependencies (Craigmile et al., 2010b,a). For this reason we examined time series plots of the RT data. Because we deleted the buffer items it is difficult to assess the extent of sequential dependencies. We have, however, ample information to assess nonstationarity. The left panels of Figure 3 show the RT series for four representative subjects. Each series is subdivided into blocks of 12 responses each, the RTs for each test trial.

The first characteristic of the RTs is that the RT distribution is not stationary over blocks. Subject 9, for example, has blocks for which all the RTs are much faster that one would expect, and Subject 14 shows less variability in the second half of the experiment than in the first. The second characteristic of the RTs is the existence of very fast or very slow responses compared to the other RTs within a block. While some models of response selection can account for fast or slow "outliers" through variability in information accumulation (e.g. Ratcliff et al., 1999; Van Zandt et al., 2000), these very fast and slow responses are too far from the mean to be explained by a single process (Craigmile et al., 2010a). Very fast responses, which are often associated with mistakes, could be associated with rapid guessing. Very slow or delayed responses could be associated with distractions or attentional failures. To avoid attributing specific characteristics to these outliers, apart from their magnitude, we will later refer to them as sub-cognitive and supra-cognitive, respectively (see Section 4.5).

3.3 RT versus number of exposures

If we model the memory process as a race between traces, the times to respond "old" to studied pictures will follow the distribution of a minimum statistic. If a trace is laid down for each exposure, the quantiles of the RT distribution should decrease as the number of study exposures increases. The right panel of Figure 3 shows boxplots of the RT distributions with respect to the number of exposures for the same four subjects as above. A decrease in the RTs with an increasing number of exposures is shown for the performing ($d' > 0.3$) Subject 17. A generally decreasing trend was observed in 6 of the other 11 performing subjects. Such a trend was not observed for any of the underperforming ($d' \leq 0.3$) subjects. These figures indicate that a model of the cognitive process should have flexibility to permit a range of effects of exposure, and suggest that traces are not laid down for each exposure.

4 Modeling

A convenient representation for the dynamics of the retrieval of a trace is based on accumulation of discrete bits of information modeled as events in a Poisson process (Audley and Pike, 1965; Pike, 1973; Van Zandt et al., 2000), which implies that the time for a trace to be retrieved follows a gamma distribution. A standard result in extreme value theory states that, after appropriate centering and scaling, the minimum of an increasing number of independent gamma random variables (RVs) converges in distribution to a Weibull RV (Fisher and Tippett, 1928; Gnedenko, 1943). Because of this, the Weibull distribution is used in race models to approximate the RT distribution when the number of racers is sufficiently large (Logan, 1988, 1992; Rouder et al., 2003). However, using the Weibull distribution to approximate the minimum gamma distribution has several drawbacks. First, when the number of racers m is small the approximation can be poor. In particular, Craigmile et al. (2013) show that, for small m, the approximation is only valid for small values of the gamma shape parameter. Second, the race model does not provide any justification for the centering and scaling required for the asymptotic result to hold (Colonius, 1995).

We use the minimum gamma race model to describe the RT distribution and fit a race model within a hierarchical Bayes framework. This approach allows us to estimate directly the number of racers in the decision process for our recognition memory experiment and, in particular, examine the probabilities with which traces are established as a function of the number of stimulus exposures. We demonstrate in Section 6 that the minimum gamma race model fits the data better than a Weibull approximation.

4.1 Race model for the experiment

Figure 4 gives a graphical representation of the hierarchical model that we constructed. In a recognition task, a subject may respond old or new to a new (unstudied) item and old or new to an old (studied) item. Given a stimulus type, old or new, our model postulates that the response is determined by a race between the old and new response processes. When the stimulus is old, the old response is itself determined by a race to retrieve any traces that may have been laid down during study. The finishing times for old and new response processes when a stimulus is new and the finishing time for the new response process when a stimulus is old are assumed to follow gamma distributions with shape and scale parameters that do not depend on the number of exposures during study. The number of exposures only influences the process that leads to a correct recognition of an old picture.

For a number E of exposures, a random number M of traces is laid down, with $0 \leq M \leq E$. During test, a race takes place between M trace retrieval processes. The winner of the race determines an overall trace process finishing time. To permit old responses to old pictures for which no traces have been laid down, there is also an independent, underlying process similar to the process that leads to an old response to a new picture. The faster of the trace process and underlying process finishing times determines the finishing time for an old response to an old stimulus. If this finishing time is shorter than the finishing time for the process that produces a new response to an old stimulus, then an old response is made. Otherwise, a new response is made.

As with the finishing times discussed above, we assume that the underlying old response process and the retrieval times for each trace are conditionally independent and identically distributed gamma RVs with shape and scale parameters that do not depend on the number of exposures. The shape and scale parameters might represent the amount of evidence necessary for a particular response and the quality of the evidence toward that response, respectively (Van Zandt et al., 2000). While we assume that the same amount of evidence is required for all components of the response process, it makes sense then that the shape parameter could vary as a function of familiarity and engagement with the task. For this reason, in Section 4.4, we allow the shape parameters to vary over block. By contrast, the scale parameters do not vary over block but may be different for each component of the response process. In particular, the old response process should be faster for old pictures while the new response process should be faster for new pictures, a characteristic that will be incorporated into the prior location parameters.

4.2 Race model with an unknown number of traces

Our analysis of the pilot data showed that the assumption that each stimulus exposure leads to a new trace is a source of lack of fit in the minimum gamma model. For this reason, we introduced a probabilistic mechanism to describe how traces are laid down following stimulus exposure. More specifically, given the number of exposures, $E = x$, for $x = 1, 2, 3, 4$, we built a model for the number of traces $M = m$ by specifying

$$\Pr(M = m|E = x, \mathbf{p}), \text{ for } m = 0, \ldots, x. \tag{1}$$

These conditional probability distributions are parameterized by the vector of conditional probabilities $\mathbf{p} = (p_0, p_1, p_2, p_3)$, where p_l is the probability of adding a trace when the number of existing traces is l. The conditional probabilities in (1) can be expressed recursively by noting that $\Pr(M = 1|E = 1, \mathbf{p}) = p_0 = 1 - \Pr(M = 0|E = 1, \mathbf{p})$ and

$$\Pr(M = m|E = x, \mathbf{p}) = p_{m-1}\Pr(M = m - 1|E = x - 1, \mathbf{p}) + (1 - p_m)\Pr(M = m|E = x - 1, \mathbf{p}),$$

where $p_m = 0$ for $m < 0$ and $\Pr(M = m|E = x) = 0$ for $m > x$.

If $M = m$ traces have been laid down during study, the trace process finishing time, V, follows a minimum gamma distribution with m components. The density of V given $M = m$, $f(v|\alpha, \beta^{A_{\text{old}}}, M = m)$, is that of a minimum gamma distribution for m independent and identically distributed racing traces, each with shape parameter α and scale parameter $\beta^{A_{\text{old}}}$. Then, assuming E exposures, marginalizing the joint distribution of V and M with respect to M yields the following marginal density for V:

$$g(v|\alpha,\beta^{A_{\text{old}}},E,\mathbf{p}) = \sum_{m=1}^{E} f(v|\alpha,\beta^{A_{\text{old}}},M=m)P(M=m|E,\mathbf{p}). \tag{2}$$

We account for subject-specific effects by constructing a hierarchy on $\alpha, \beta^{A_{\text{old}}}$, and $\mathbf{p}$. The distributions of α and $\beta^{A_{\text{old}}}$ and their theoretical interpretations are described in the following sections. For Subject k, we assume that the elements of $\mathbf{p_k} = (p_{0,k}, p_{1,k}, p_{2,k}, p_{3,k})$ are conditionally independent with distributions

$$\text{logit}(p_{l,k}) \sim N(\mu_{p_l}, \sigma^2_{p_l}), \text{ for } l = 0, 1, 2, 3. \tag{3}$$

4.3 Stimulus effects

The finishing time O^{old} for an old response to an old stimulus is modeled as

$$O^{\text{old}} = \min(U, V), \tag{4}$$

where U denotes the underlying old-to-old process finishing time with

$$U \sim \text{Gamma}\left(\alpha, \beta^{O_{\text{old}}}\right) \tag{5}$$

and V denotes the independent trace process finishing time which has an unknown number of racers and density with parameters α and $\beta^{A_{\text{old}}}$ given in (2). The other three finishing times, for responses new to old, old to new and new to new are denoted by N^{old}, O^{new}, and N^{new}, respectively. They are assumed to be distributed as follows:

$$N^{\text{old}} \sim \text{Gamma}\left(\alpha, \beta^{N_{\text{old}}}\right);\ O^{\text{new}} \sim \text{Gamma}\left(\alpha, \beta^{O_{\text{new}}}\right);\ \text{and } N^{\text{new}} \sim \text{Gamma}\left(\alpha, \beta^{N_{\text{new}}}\right). \tag{6}$$

We denote the overall finishing time for each stimulus/response combination by C and introduce the response $R = 0$ or 1 for new or old responses, respectively. Then we can write

$$R|O^{\text{st}}, N^{\text{st}} \ = \ I\left(O^{\text{st}} \leq N^{\text{st}}\right), \quad \text{and} \quad C|R, O^{\text{st}}, N^{\text{st}} \ = \ O^{\text{st}}R + N^{\text{st}}(1-R), \tag{7}$$

where the stimulus st = old or new. In the construction of the hierarchical model, the distributions of the four finishing times (and hence those of C and R) will be allowed to vary with subject $k = 1, \ldots, S$, block $j = 1, \ldots, B$, and trial $i = 1, \ldots, n$. Correspondingly, appropriate subscripts will be used when needed.

Finally, denoting by $\boldsymbol{\beta}_k$ the vector of scale parameters $\left(\beta_k^{A_{\text{old}}}, \beta_k^{O_{\text{old}}}, \beta_k^{N_{\text{old}}}, \beta_k^{O_{\text{new}}}, \beta_k^{N_{\text{new}}}\right)$ for subject k, we assume that each component, independently, has a log normal distribution:

$$\log(\beta_k^c) \sim N(\mu_{\beta^c}, \sigma^2_{\beta^c}), \quad c = A_{\text{old}}, O_{\text{old}}, N_{\text{old}}, O_{\text{new}}, N_{\text{new}}. \tag{8}$$

4.4 Time-varying elements

Previous work demonstrated the importance of modeling RT data as arising from processes that include trends and local dependencies (Craigmile et al., 2010b) to account for subjects' varying levels of attention, fatigue and increasing familiarity with the task. Our preliminary analysis of the pilot data showed poor fits of a reduced model that did not include trend and local dependence and our final analysis captures these components of variation by way of a model for the shape parameter α of the gamma distributions, which represents the information thresholds of the cognitive process. For each subject k we assume the logarithm of $\boldsymbol{\alpha}_k = (\alpha_{1k}, \ldots, \alpha_{20\,k})$ varies over block as an autoregressive process of order one, AR(1), plus a linear trend over blocks. Conditional on the initial mean $\mu_{\alpha,k}$, the slope for the linear trend $s_{\alpha,k}$, the overall standard deviation $\sigma_{\alpha,k}$, and the autoregressive parameter $\phi_{\alpha,k}$, the likelihood of $\boldsymbol{\alpha}_k$ for each subject k is described by

$$\begin{aligned} \log \alpha_{1,k} &\sim N\Big(\mu_{\alpha,k} + s_{\alpha,k}, \sigma^2_{\alpha,k}/(1 - \phi^2_{\alpha,k})\Big); \\ \log \alpha_{j,k} | \log \alpha_{j-1,k} &\sim N\Big(\mu_{\alpha,k} + j s_{\alpha,k} + \phi_{\alpha,k}\left[\log \alpha_{j-1,k} - \mu_{\alpha,k} - (j-1) s_{\alpha,k}\right], \sigma^2_{\alpha,k}\Big), \\ &\qquad j = 2, \ldots, B. \end{aligned} \tag{9}$$

4.5 Ancillary processes

Preliminary, unsatisfactory fits of a minimum gamma model indicated that the finishing time for the decision process of interest cannot be observed directly because there are other processes that affect an observed RT and response. Our analysis of the pilot data, along with previous research by Craigmile et al. (2010b), suggested three such ancillary processes. The first process accounts for residual processes that are not related to making a decision, such as perceptual encoding. This process adds to the time for the decision process. Two additional processes, incorporated into an overall mixture model, account for responses that are either too fast (sub-cognitive) or too slow (supra-cognitive). The importance of including such components in the modeling of recognition memory performance has been demonstrated by Province and Rouder (2012). They provided evidence for the existence of mixtures of different processes in recognition memory data and showed that increasing the number of exposures of a studied item results in changing the probabilities that a response is based on the result of these sub- or supra-cognitive processes (guessing, in particular). These component processes will be described separately in Sections 4.5.1 and 4.5.2 and the overall mixture process that determines which process is active at any given time will be described in Section 4.6.

4.5.1 The residual time

We use T^0 to represent the residual time taken by encoding and/or motor processes. This residual time is added to the decision time. One way to model the effects of encoding and motor processes is to subtract a fixed, subject-specific value from the RTs, where the value is a deterministic function (such as a certain fraction) of the fastest observed RT (Peruggia et al., 2002; Logan et al., 2014). However, treating T^0 as a function of the minimum RT fails to consider that some of the fast responses in the data (including the minimum RT) might be unrelated to the cognitive and encoding processes and therefore too fast to be used as an estimate of T^0. So, we treat T^0 as a subject-specific random variable and allow a different model component to account partially for these fast responses. For subject $k = 1, \ldots, S$, we assume

$$T_k^0 \sim N^+(\mu_{T^0}, \sigma_{T^0}^2), \tag{10}$$

where $N^+(\cdot, \cdot)$ denotes a normal distribution truncated to the positive half-line. Using our pilot data, we specified a prior for $\{\mu_{T^0}, \sigma_{T^0}^2\}$ that kept T_k^0 within a reasonable range.

4.5.2 Sub-cognitive and supra-cognitive responses

For subject $k = 1, \ldots, S$, block $j = 1, \ldots, B$, and trial $i = 1, \ldots, n$, we modeled sub-cognitive time G and supra-cognitive time H with log-normal distributions:

$$\log(G_{ijk}) \sim N(\mu_{G,k}, \sigma^2_{G,k} = 1), \quad \text{and} \quad \log(H_{ijk}) \sim N(\mu_{H,k}, \sigma^2_{H,k} = 0.4^2). \tag{11}$$

As explained in Section 4.7, we specified informative prior distributions for μ_G and μ_H to ensure that, with high probability, G and H take on appropriate values.

4.6 Full likelihood

Putting all components together, the model for the observed RTs T is a mixture of sub-cognitive times G, decision process times C, and supra-cognitive times H. For responses generated by the decision process or the supra-cognitive process, the observed time includes the non-decision process T^0. T^0 is not added to the sub-cognitive times because those times were generally too small to have been generated by any encoding processes. A priori, we assume that G and H have small probabilities of determining responses and that they are conditionally independent and independent of the experimental conditions.

The mixture probabilities of the three processes, which vary over blocks j and subjects k, are denoted by $\mathbf{q}_{jk} = (q^G_{jk}, q^C_{jk}, q^H_{jk})$, with q^G_{jk}, q^C_{jk}, $q^H_{jk} \in (0, 1)$, and $q^G_{jk} + q^C_{jk} + q^H_{jk} = 1$. Then, for $i = 1, \ldots, n$ and given $\mathbf{q}_{jk}$, we assume that ξ_{ijk} are independent draws from a discrete distribution $f_\xi(\xi_{ijk}|\mathbf{q}_{jk})$ taking on values in $\{0, 1, 2\}$ with probabilities $\mathbf{q}_{jk}$, and that, given ξ_{ijk}, G_{ijk}, C_{ijk}, H_{ijk}, and T^0_k,

$$T_{ijk} = I(\xi_{ijk} = 0)G_{ijk} + I(\xi_{ijk} = 1)(C_{ijk} + T^0_k) + I(\xi_{ijk} = 2)(H_{ijk} + T^0_k). \tag{12}$$

Then the likelihood function of our model is

$$\begin{aligned}
&\left(\prod_{k=1}^{S} f_p(\mathbf{p}_k|\boldsymbol{\mu}_p,\boldsymbol{\sigma}_p)\, f_\alpha(\boldsymbol{\alpha}_k|\mu_{\alpha,k},\sigma_{\alpha,k},\phi_{\alpha,k})\, f_\beta(\boldsymbol{\beta}_k|\{\mu_{\beta^c}\},\{\sigma_{\beta^c}\})\, f_{T^0}(T_k^0|\mu_{T^0,k},\sigma_{T^0,k})\right) \\
&\prod_{k=1}^{S}\prod_{j=1}^{B}\Bigg(\prod_{i\in\{\xi_{ijk}=0\}} f_G(T_{ijk}|\mu_{G,k},\sigma_{G,k}) \prod_{i\in\{\xi_{ijk}=2\}} f_H(T_{ijk}-T_k^0|T_k^0,\mu_{H,k},\sigma_{H,k}) \\
&\qquad \prod_{i\in\{\xi_{ijk}=1\}} f_{C,R}(T_{ijk}-T_k^0,R_{ijk}|T_k^0,E_{ijk},\mathbf{p}_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k) \prod_{i=1}^{n} f_\xi(\xi_{ijk}|\mathbf{q}_{jk})\Bigg).
\end{aligned} \tag{13}$$

The density functions in (13) are determined by the structure we have discussed so far. Specifically, f_p comes from (3), f_α from (9), f_β from (8), f_{T^0} from (10), f_G and f_H from (11), and $f_{C,R}$ from the relations and distributions given in (2), and (4)–(7).

4.7 Prior settings

Our prior specifications strike a balance between computational ease and model realism. Whenever possible, our priors, summarized in the supplemental material, are informative and incorporate knowledge about parameters derived from the pilot data.

We specified the prior distributions of the mixture components G, C, and H to promote a stochastic ordering $\prec$ among them so that $G \prec C + T^0 \prec H + T^0$. The distribution of C is influenced by the assumptions about $\mu_{\alpha,k}$, the prior mass of which is concentrated away from zero to ensure that C and G can be distinguished. The prior mass of $\log(\sigma_{\alpha,k})$ concentrates on small values to control the magnitude of the changes of α over blocks and we assumed a normal prior with large variance for the slope parameter $s_{\alpha,k}$. The prior ordering $G \prec C + T^0$ is promoted by the fact that the prior parameters for the distribution of G favor very fast times and by the interplay between G and T^0: all responses faster than T^0 are sub-cognitive, G, with probability one. The prior parameters for the distribution of H favor very slow times compared to the typical times for the cognitive decision process, C, which promotes the prior ordering $C + T^0 \prec H + T^0$.

We used the pilot data to perform preliminary diagnostics, sensitivity analyses, and model refinement on closely-related precursors of our final hierarchical model. A comparison between the separate fits to each subject's data and the hierarchical fits did not reveal any inconsistencies in the estimated posteriors. We also determined that, with the notable exception of **p**, the posterior distributions of the parameters of the decision process are not very sensitive to moderate changes in the prior specifications. The specification of the prior for **p** is inextricably tied to the specification of a model for the trace process finishing time V described in Section 4.2. This led us to examining a number of model structures for the trace process, including the possibilities that all p_l may be different, that $p_0 = 1$, and that $p_2 = p_3$. The assumption that $p_0 = 1$ corresponds to believing that a trace is always laid down when a stimulus is exposed for the first time, which avoids the necessity of including a process that can produce old responses to old stimuli even in the absence of any traces. The assumption that $p_2 = p_3$ corresponds to believing that, from a certain point on, the probability of adding a new trace becomes constant (as suggested by Figure 5). The hyperparameters of the prior distributions of the p_ls for each of these cases is reported in the supplemental material. In the analyses that follow we make use of the default "base" model, with a choice of prior for the p_ls centered at 0.5 and rather spread out. This prior choice for p_l ignores the information from the preliminary study (because such information was based on a slightly different model) and tries to make sure that the posterior distributions are driven by the data. We discuss the sensitivity of the priors distributions for p_l in Sections 5.1 and 6 below.

5 Posterior inferences

We fit our models using the MCMC algorithm described in the supplemental material. For each fitted model we ran four independent MCMC chains that started from random initial conditions. For each chain we discarded a burn-in of 1,000 samples and obtained 6,000 draws (a subsampling of 120,000 further draws, thinned every 20). Thus, our posterior inferences are based on 24,000 draws from the posterior. We confirmed convergence of the MCMC chains by inspecting the trace plots and computing Gelman and Rubin's multiple sequence diagnostic (Brooks and Gelman, 1998). We discuss the findings from our analyses based on draws from the posterior distributions of the model parameters.

5.1 Exploring the trace mechanism

Conditional on the data, the posterior distributions of p_l describe the probability of adding a trace given that the existing number of traces in memory is equal to l. Figure 5 presents, for two different priors, boxplots of the posterior predictive draws of p_l computed from the posterior MCMC samples of μ_{p_l} and $\sigma^2_{p_l}$ for each $l = 0, 1, 2, 3$. Panel (a) shows the posterior predictive draws for our model in which we use a prior for p_l centered at 0.5. The boxplots show that the conditional probabilities of laying down traces are widely variable and that most of the mass of each conditional distribution is concentrated away from one. The posteriors of p_0 and p_1 are closer to zero and slightly more spread out than the prior. The posteriors of p_2 and p_3 are roughly centered at the prior mean, but are much more spread out than the prior. Given the degree of uncertainty in the posteriors of p_2 and p_3, we investigated the influence of the prior by moving the prior mean away from 0.5. The posteriors of p_0 and p_1 were not strongly influenced by the location of the prior, but the posteriors of p_2 and p_3 were pulled in the direction of the prior and were quite spread out.

An extreme case is one in which the prior mass for p_l is highly concentrated near 1. This is very close to a prior that puts a point mass at one, which is consistent with models that assume a trace is laid down deterministically with each exposure. However, this prior is fully supported on $(0, 1)$, allowing us to see whether the data dispel this belief. Panel (b) of Figure 5 displays the posterior predictive draws for this case. The marked downward shift of the posteriors for p_0 and p_1 compared to the prior gives strong evidence that deterministic laying down of traces is not supported by the data. Perturbations to the variability of the prior centered at 0.5 had little effect on the posteriors.

These posterior summaries demonstrate that the actual number of traces M laid down with repeated stimulus exposures is very likely to be less than the number of exposures E. Therefore, our model provides a mechanism for evaluating the number of traces established during study and the contribution of these traces to memory performance.

5.2 Identifying ancillary processes

The postulated mixture model of (12) provides a stochastic mechanism, characterized by ξ_{ijk}, to assign an RT to the regular race process C or to one of the two ancillary processes G and H, while accommodating the residual time T^0. One way to characterize subjects as performing or not is to examine the proportion of RTs that come from ancillary processes. We can contrast the extents and locations of the regions over which there is uncertainty about the allocation of an outcome to different model components with the densities of the observed RTs over these regions. These contrasts give an indication of how well the model was able to discriminate between outcomes arising from the cognitive process versus the ancillary processes. We expect to see the posterior probability that an RT arises from the sub-cognitive process to decrease as RT increases, whereas we expect to see the posterior probability that an RT arises from the supra-cognitive process to increase as RT increases.

For illustration, we continue to look at Subjects 9, 11, 14, and 17. For each subject, Figure 6 plots the posterior probabilities that the RT on a given trial is the result of either sub-cognitive (circles) or supra-cognitive (triangles) processing against the observed RT. The vertical line shows the posterior median of the residual times T^0, and the vertical dashed lines indicate 95% posterior credible intervals for T^0. RTs below T^0 are always identified as coming from the sub-cognitive process. The short vertical line segments plotted at the bottom of each panel correspond to the observed RTs. Ranges of the time axis with higher densities of observed RTs are rendered with darker shading.

Consider first the sub-cognitive process G. For Subject 17, the second-highest performing subject, the estimated allocation probability to the sub-cognitive process is essentially zero (below 0.005) for all outcomes. For Subject 11, 98.7% of the allocation probabilities to the sub-cognitive process are either below 0.01 or above 0.99. For Subject 14, 97.1% of these probabilities are either below 0.01 or above 0.99. By contrast, for Subject 9, one of the lowest-performing subjects, there are many observed outcomes (35.5%) whose allocation probabilities to the sub-cognitive process span the range from 0.01 to 0.99. For this subject, the RTs for the outcomes that are not clearly identified as belonging to the sub-cognitive or the cognitive process are not concentrated near the mean of the residual time T^0 but extend considerably to its right. In addition, in contrast to the other subjects' posteriors, the posterior uncertainty about T^0 is much greater.

The connection between performance and the behavior of the estimated allocation probabilities to the sub-cognitive process is confirmed by the analysis of the remaining subjects. The plots for four of the seven underperforming subjects, as well as the plot for the marginally performing Subject 19 with a d' score of 0.305, also resembled the plot for Subject 9.

Considering now the supra-cognitive process H, we look at Subject 14. The estimated allocation probabilities to the slower process for Subject 14 rise in the tail of the RT distribution. However, these probabilities are not strictly increasing: slower observed RTs may be classified as arising from the H component with smaller posterior probabilities than other faster RTs. This is because the mixture probabilities $\mathbf{q}_{jk}$ change over block to account for fluctuations over time in the relations between the cognitive and the ancillary processes.

These results indicate that our model can differentiate between performing subjects with regular patterns of behavior and more erratic, underperforming subjects. Furthermore, our model can provide an explanation of poor performance in terms of different processes (sub-cognitive and supra-cognitive) that detract from the subjects' abilities to perform the task, as well as in terms of failures of the cognitive process itself. In the next section, we characterize subject performance in terms of the cognitive process.

5.3 Measures of performance

Subjects in our experiment performed the recognition task with varying degrees of success, as indexed by d'. We investigated three additional complementary measures of performance. First, we examined how effectively the subjects encoded the pictures in terms of the individual-level probability of adding a first trace, $p_{0,k}$. The larger this probability, the more effectively traces were laid down. Second, we examined how well subjects could extract information from a test picture by computing the ratio of the new and old accumulation process scale parameters when an old item was presented, $\lambda_k = \beta_k^{N_{\text{old}}}/\beta_k^{A_{\text{old}}}$. If a picture is old, it should be easier to extract information from it that is consistent with an old response as opposed to a new response, so the ratio λ_k should be larger than one. Third, we computed the proportion of non-sub-cognitive responses ρ_k (the proportion of trials for which $\xi_{ijk} \neq 0$) in the fastest 10% of RTs. Figure 7 illustrates the relationship between d' and the three statistics $p_{0,k}$, λ_k and ρ_k. If a subject is performing the task well and is able to remember the pictures that he or she studied, we would expect to see that subject's statistics located at approximately the locations shown by Subject 15 in Figure 7: high values of $p_{0,k}$, λ_k and ρ_k.

The cluster of subjects that appear consistently below the dotted line indicating $d' = 0.3$ fail to perform the task well, but for different reasons. For example, Subject 6 does not have a high number of sub-cognitive responses (the mean of ρ_6 is high), and, when a trace is present, is able to extract relevant information from memory (the mean of λ_6 is high). However, Subject 6 has difficulty laying down traces (the mean of $p_{0,6}$ is not very high). By contrast, Subject 20 is better at laying down traces (the mean of $p_{0,20}$ is higher) and can extract relevant information from those traces (the mean of λ_{20} is over 1.0), but executes a very large number of sub-cognitive responses (the mean of ρ_{20} is less than 0.2).

We can see from Figure 7 that subjects who perform well also do so for different reasons. Subject 10 has difficulty laying down traces (this subject has the smallest mean value of $p_{0,k}$) but is able to extract a great deal of information from these traces (the mean of λ_{10} is 2.0). By contrast, Subject 17 does not extract as much information from stored traces (the mean of λ_{17} is roughly half the mean of λ_{10}), but is better able to lay down traces in the first place (the mean of $p_{0,17}$ is more than twice the mean of $p_{0,10}$). Most of the performing subjects are consistent, however, in showing low frequencies of sub-cognitive responses, higher probabilities of laying down traces, and high extraction rates.

While we formulated the three statistics $p_{0,k}$, λ_k and ρ_k somewhat arbitrarily, their analysis highlights the advantages of modeling processes that are ancillary to the cognitive process of interest. These analyses also demonstrate the utility of modeling all the data, instead of discarding data that are difficult to model or that do not fit into our conceptions of the process of interest. Most importantly, we can identify the different ways that subjects succeed or fail at performing a task, identification that is impossible without a model structure incorporating ancillary processes.

6 Model validation and model comparison

We now use out-of-sample predictive diagnostics to validate the minimum gamma model and to compare it to an alternative Weibull model. We chose to perform out-of-sample prediction because it is a challenging standard for model validation and comparison that tends to penalize overfitting of the data. Both in the validation exercise of Section 6.1 and in the model comparison of Section 6.3 we computed fits conditional on the observations in the first 18 blocks and predicted the held-out observations in blocks 19 and 20. The choice of predicting the last several blocks in the sequence is natural because of the sequential nature of the data. Using one-tenth of the data to construct the hold-out sample strikes a balance between the size of the training and testing data sets, and is consistent with common recommendations in cross-validation (Hastie et al., 2009). This choice allowed us to predict a sizable portion of the observations while retaining the ability to reliably estimate the predictive densities.

6.1 Validation via predictive distributions for future blocks

To validate the minimum gamma model we estimated the posterior predictive distributions of the three mixture components in Equation (12) conditional on the observations in the first 18 blocks, and determined how well the three mixture components can, in concert, capture the variability among the held-out observations in blocks 19 and 20.

Let $\theta^{(1)}, \ldots, \theta^{(L)}$ denote the MCMC draws of the minimum gamma model parameters conditional on the observations in the first 18 blocks. (Because there are no posterior MCMC samples available for the block-specific parameters α_{jk}, $j = 19, 20$, each parameter vector $\theta^{(\ell)}$ includes values for $\alpha^{(\ell)}_{19,k}$ and $\alpha^{(\ell)}_{20,k}$ generated from the posterior MCMC samples for the previous blocks and the AR(1) model of Section 4.4.) For each $\theta^{(\ell)}$ and each of blocks 19 and 20, we generated one observation from seven posterior predictive distributions: one observation from the distribution of the sub-cognitive times G, one observation from the distribution of the supra-cognitive plus non-decision process times $H + T^0$, and, for each number of exposures from 0 to 4, one observation from the distribution of the decision process plus non-decision process times $C + T^0$.

For the four subjects shown in Figure 3, Figure 8 displays (on a logarithmic scale) summaries of the empirical predictive distributions based on the MCMC draws together with the observed RTs (shown as crosses). In each panel, the five boxplots summarize the predictive RT distributions of the cognitive process for the corresponding number of exposures. The two dashed lines at the top of each panel delineate the interquartile range of the predictive RT distributions for the supra-cognitive process and the two dashed lines at the bottom of each panel delineate the interquartile range of the predictive RT distributions for the sub-cognitive process. In accordance with our distributional assumptions, the three predictive distributions concentrate most of their masses on different regions of the RT range. As a diagnostic check, we want to assess if the held-out observations in blocks 19 and 20 can be well explained by (at least) one of the three mixture components. This was the case for the subjects in the figure as well as for the other 16 subjects whose predictive distributions are not shown. We note that, while many observations are well explained by the cognitive process, the two other components of the mixture are needed to account for some of the more extreme observations.

For example, the observation corresponding to Subject 14, block 20, and 2 exposures is very likely to have arisen from the supra-cognitive process. Similarly, some of the observations for the subjects not included in the figure are very likely to have arisen from the supra-cognitive process. Note also that, for Subject 9, there is often uncertainty about the processes responsible for producing some of the observations, as many observations fall in regions where two processes have overlapping mass (e.g., blocks 19 and 20, 0 exposures). This is consistent with the posterior probabilities presented in Figure 6. Additional model diagnostics performed on the entire data set, including the calculation of conditional predictive ordinates (Geisser and Eddy, 1979) for the detection of outliers did not raise any serious lack-of-fit concerns.

6.2 A competing Weibull model

As we discussed in Section 4, Logan and colleagues described the finishing times of trace-race models with Weibull distributions (Logan, 1988, 1992). It is important to note that, because of the small number of traces that could be established in our experimental task, our race model is not appropriate for the kinds of skill-acquisition data that Logan and colleagues have modeled with the Weibull distribution, but the Weibull has been proposed as a general model for describing RT data (Rouder et al., 2003).

We compare our proposed minimum gamma model to an approximating Weibull model. To establish as close a correspondence as possible between the various elements of the two models, we simply replaced the minimum gamma components of the model determining the processing times O^{old}, N^{old}, O^{new} and N^{new} (see Figure 4) with Weibull random variables. Consistent with previous applications of the Weibull model, we assumed that the number of exposures E affects the scale parameter in the Weibull distribution of O^{old} according to a power law, so that $O^{\text{old}}_{ijk}|E \sim \text{Weibull}(\alpha^{W}_{jk}, \beta^{W,O_{\text{old}}} E^{-\eta^{W,O_{\text{old}}}})$, where the superscript W distinguishes the shape $\alpha^{W}_{jk} > 0$ and scale $\beta^{W,O_{\text{old}}} > 0$ parameters of the Weibull distribution from those of the minimum gamma model. The additional parameter $\eta^{W,O_{\text{old}}} > 0$ determines the rate of reduction in RT as E increases. The Weibull distributions for the finishing times N^{old}, O^{new} and N^{new} have scale parameters $\beta^{W,N_{\text{old}}}$, $\beta^{W,O_{\text{new}}}$ and $\beta^{W,N_{\text{new}}}$, and a common shape parameter α^{W}_{jk} which is also shared by the distribution of O^{old}.

While the Weibull approximation to the minimum gamma distribution may be accurate given a sufficient number of racing traces, it is not easy to relate the Weibull parameters directly to the theoretically-interpretable minimum gamma model parameters. This makes the interpretation of the Weibull parameters difficult and complicates the specification of the priors. To make the comparison between the two models as fair as possible, we selected informative priors for the Weibull model based on preliminary fits of the pilot data (see the supplement). We followed the same procedures that we performed for the minimum gamma model, including similar MCMC convergence and model validation checks. analyses. The posterior distributions were similar for those parameters that have a common theoretical interpretation in the two models. These similarities concerned not only the shape and scale parameters of the gamma and Weibull distributions, but also ancillary parameters such as T^0 and ξ.

6.3 Comparison via marginal likelihoods for future blocks

To evaluate the two alternative models, we performed out-of-sample predictive evaluations similar to the validation diagnostics of Section 6.1. This involved fitting the first 18 blocks of RTs and responses for all subjects and measuring prediction accuracy through the marginal likelihoods of the hold-out observations in blocks 19 and 20. More specifically, let $m = 1$ denote the minimum gamma model and let $m = 2$ denote the Weibull model. Also, let $\boldsymbol{D}_j$ denote the observations (data) in the jth block for *all* subjects. The data $\boldsymbol{D}_j$ include both the RTs and the decisions. The marginal likelihood of $(\boldsymbol{D}_{19}, \boldsymbol{D}_{20})$ under model m is

$$f_m(\boldsymbol{D}_{19}, \boldsymbol{D}_{20}|\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{18}) = \int f_m(\boldsymbol{D}_{19}, \boldsymbol{D}_{20}|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m|\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{18})d\boldsymbol{\theta}_m, \tag{14}$$

where $\pi_m(\boldsymbol{\theta}_m|\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{18})$ is the posterior distribution of the parameters $\boldsymbol{\theta}_m$ for model m, given $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{18}$. The two models can be compared via the ratio of their marginal likelihoods, either on the natural scale, $f_1(\boldsymbol{D})/f_2(\boldsymbol{D})$, or on the logarithmic scale, $\log f_1(\boldsymbol{D}) - \log f_2(\boldsymbol{D})$. This ratio of marginal likelihoods can be thought of as the Bayes factor (Jeffreys, 1967) for the data in blocks 19 and 20, using as the prior distribution for $\boldsymbol{\theta}_m$ the posterior distribution $\pi_m(\boldsymbol{\theta}_m|\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{18})$. There are no closed-form expressions for the marginal likelihoods for our complex models, and so we estimated (14) via Monte Carlo. The standard errors of these estimates were obtained using a moving block bootstrap (e.g. Lahiri, 2013), thus duly accounting for the autocorrelations in the MCMC samples.

For both models, the parameter vectors $\boldsymbol{\theta}$ include generated values of the block-specific parameters α and q. Specifically, $\boldsymbol{\theta}_1^{(\ell)}$ includes values of $\alpha_{jk}^{(\ell)}$ and $q_{jk}^{(\ell)}$, and $\boldsymbol{\theta}_2^{(\ell)}$ includes values of $\alpha_{jk}^{W(\ell)}$ and $q_{jk}^{W(\ell)}$, $j = 19, 20$. There are no posterior MCMC samples for these parameters but they are needed to compute the Monte Carlo estimate of (14). As in Section 6.1, the α values were generated using the posterior MCMC samples for the previous blocks and the AR(1) model of Section 4.4. The q values are not tied to any of the parameters in the previous blocks, and so we set them equal to their prior means.

Figure 9 displays the increase (crosses) in the estimated marginal log-likelihood of the Weibull model, relative to the minimum gamma (base) model for each subject. The vertical gray lines indicate the 95% confidence intervals of the estimated increase derived from the moving block boostrap. Overall, the minimum gamma model is better than or comparable to the Weibull model for all subjects, with the exception of Subject 9. The same figure also presents marginal log-likelihood comparisons between the minimum gamma base model and the three other minimum gamma models mentioned in Section 4.7. The first of these models has $p_0 = 1$ (plus signs), the second has $p_2 = p_3$ (circles), and the third has priors for p_l concentrated near 1. Compared to the base model, the performance of the first and third models was worse for the majority of subjects, which agrees with the conclusion reached in Section 5.1. However, we could see no substantive differences between the base model and the $p_2 = p_3$ model across subjects. Choosing the model with less constraints, we used the base model to report our inferential findings throughout this article.

As we noted above, comparing marginal likelihoods after conditioning on a substantial amount of data promotes stability in the estimates of the likelihood differences. Using the posteriors of the parameters conditional on the first 18 blocks of data promotes comparisons that are more fair. That is, because the parameterizations of the two models are not entirely comparable, the calculation of a Bayes factor for the entire data set might be unduly influenced by idiosyncrasies in how we specified the priors. This concern, however, is mitigated by the preliminary updating which causes the distributions on the model parameters to concentrate on regions that are strongly supported by the data.

7 Discussion

In this article, we have presented a model of memory retrieval that assumes that traces for studied items race against each other to produce a recognition response. In addition, these traces race against other possible responses. This structure can produce both correct and incorrect responses to any stimulus as well as the time to make them. Similar models have been proposed to explain recognition memory performance in the past, but it is important to stress that our model is not intended as a competitor to these earlier models or to more recent, theoretically and empirically motivated models of memory. The purpose of our model is to explore different aspects of memory performance by establishing a quantitative framework within which the trace construction (cognitive) process operates together with extraneous sub-cognitive and supra-cognitive processes.

Construction of this more complete model has at least three benefits. First, by acknowledging that sometimes subjects do not perform the task before them, we are able to include all the experimental data in the analysis. We are not forced to eliminate trials on which a person responded too slowly, nor are we forced to remove subjects who were unable to perform the task to our satisfaction. Second, the model is able to identify and isolate these extraneous processes from the memory retrieval process of interest, so that the conclusions that we draw from the model posteriors are uncontaminated by potentially confounding influences. Third, from these posterior distributions, we are able to see that subjects perform (and fail to perform) the task in different ways, providing a much more nuanced picture of the ways that recognition memory unfolds.

We conclude from our analysis that presentation of items for study does not necessarily lead to establishing traces for those items. Trace construction is probabilistic. As discussed in Section 1, some older models assume that this process is deterministic. More recent models also assume that traces are established deterministically, but allow them to decay with time, and be error-prone or subject to interference. These corrupted traces lead to poor recognition performance later (Shiffrin and Steyvers, 1997; Dennis and Humphreys, 2001). The probability of establishing a trace may represent the uncertainty inherent in a trace due either to infidelity in the representation of a study item or a lack of a trace for that item. Note, however, that in our model (which does not incorporate trace corruption), every existing trace, whether corrupt or not, takes part in the race to respond "old." Therefore, the only way that a corrupted trace can lead to poorer recognition memory is if it is never established at all. For this reason, the probability of laying down traces in our model encompasses and oversimplifies a wide range of cognitive processes established by other, more sophisticated memory models. Further refinement of this component of our model will be necessary if it is to be taken seriously as a model of recognition memory, rather than as a tool to decompose recognition memory performance.

The most important contribution of this work is the presentation of a method for decomposing individual subject performance. By including model components to represent ancillary (non-cognitive) processes, we showed how good and bad performance can arise in different ways. Considering bad performance, is it the case that subjects are not doing the task they were asked to do, or are they trying but unable to remember the pictures they were shown? This is an important distinction, because while we might want to remove subjects from the data set who are not doing the task, we would not want to remove the subjects who were trying but could not remember what they studied. The first group of subjects do not contribute to our understanding of the processes of interest, whereas the second group does. Modeling the performance of all subjects within the same hierarchy allows us to make this distinction, and even underperforming subjects can contribute to our understanding of the process.

We found that subjects who perform the task well avoid making fast responses and are either good at laying down traces, or at extracting information from existing traces, or both. Subjects who do not perform the task well show some combination of high-frequency sub-cognitive processing, low probabilities of laying down traces or low rates of information extraction. The model lets us make precise quantitative statements about these elements of performance as illustrated in Figure 7. These quantitative statements could be used to classify people as they perform different tasks, with potential applications to latent variable modeling, individual difference research, medical diagnostics, and so on.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grants No. SES-1024709, No. SES-1424481 and No. DMS-1407604. We are indebted to the associate editor and three anonymous reviewers for their comments on earlier drafts of this article that greatly improved its quality.

References

Audley and Pike (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology*, 18:207–225.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychol. Bull.*, 74:81–99.

Bloch, D. (2007). *Aristotle on Memory and Recollection*. Brill, Leiden.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Colonius, H. (1995). The instance theory of automaticity: Why the Weibull? *Psychol. Rev.*, 102:744–750.

Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2010a). Detrending response time series. In Chow, S.-M., Ferrer, E., and Hsieh, F., editors, *Statistical Methods for Modeling Human Dynamics: An Interdisciplinary Dialogue*, volume 4. Taylor and Francis, New York, NY.

Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2010b). Hierarchical Bayes models for response time data. *Psychometrika*, 75:613–632.

Craigmile, P. F., Peruggia, M., and Zandt, T. V. (2013). A Bayesian hierarchical model for response time data providing evidence for criteria changes over time. In Edwards, M. C. and MacCallum, R. C., editors, *Current Issues in the Theory and Application of Latent Variable Models*. Taylor and Francis, New York, NY.

Dennis, S. and Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychol. Rev.*, 108:452–478.

Ebbinghaus, H. (1913). *On memory: A contribution to experimental psychology*. Teachers College, New York, NY.

Egan, J. P. (1958). Recognition memory and the operating characteristic. Technical Report AFCRC-TN-58-51, Hearing and Communication Lab., Indiana Univ., Bloomington, IN.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d’une série aléatoire. *Annals of Mathematics*, pages 423–453.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York, NY.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*. Springer, New York, NY.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychol. Rev.*, 95:528–551.

Hirshman, E. (1995). Decision processes in recognition memory: criterion shifts and the list-strength paradigm. *J. Exp. Psychol.: Learn. Mem. Cogn.*, 21:302–313.

Jeffreys, S. H. (1967). *Theory of Probability, 3rd Edition*. Clarendon Press, Oxford, UK.

Kates, W. R., Krauss, B. R., AbdulSabur, N., Colgan, D., Antshel, K. M., Higgins, A. M., and Shprintzen, R. J. (2007). The neural correlates of non-spatial working memory in velocardiofacial syndrome (22q11.2 deletion syndrome). *Neuropsychologia*, 45:2863–2873.

Lahiri, S. N. (2013). *Resampling methods for dependent data*. Springer, New York, NY.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychol. Rev.*, 95:492.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *J.Exp. Psychol.: Learn. Mem.Cogn.*, 18:883–914.

Logan, G. D., Van Zandt, T., Verbruggen, F., and Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychol. Rev.*, 121:66–95.

Murdock Jr, B. (1962). The serial position effect of free recall. *J. Exp. Psychol.*, 64:482–488.

Musen, G. and Treisman, A. (1990). Implicit and explicit memory for visual patterns. *J. Exp. Psychol.: Learn. Mem. Cogn.*, 16:127–137.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *J. Exp. Psychol.: Learn. Mem. Cogn.*, 13:87–108.

Nosofsky, R. M., Kruschke, J. K., and McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *J. Exp. Psychol.: Learn. Mem. Cogn.*, 18:211–233.

Nosofsky, R. M. and Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological review*, 104:266–300.

Peruggia, M., Van Zandt, T., and Chen, M. (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. In *Case Studies in Bayesian Statistics*, volume 6, pages 319–334. Springer, New York.

Pike, R. (1973). Response latency models for signal detection. *Psychol. Rev.*, 80:53–68.

Province, J. M. and Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, 109:14357–14362.

Raaijmakers, J. G. W. and Shiffrin, R. M. (1981). Search of associative memory. *Psychol. Rev.*, 88:93–134.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.*, 85:59–108.

Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.*, 111:333–367.

Ratcliff, R., Van Zandt, T., and McKoon, G. (1999). Comparing connectionist and diffusion models of reaction time. *Psychol. Rev.*, 106:261–300.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., and Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80:491–513.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., and Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68:589–606.

Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin and Review*, 4:145–166.

Taatgen, N. A., Huss, D., Dickison, D., and Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *J. Exp. Psychol.: General*, 137:548–565.

Turner, B. M., Van Zandt, T., and Brown, S. D. (2011). A dynamic, stimulus-driven model of signal detection. *Psychol. Rev.*, 118:583–613.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.*, 108:550–592.

Van Zandt, T., Colonius, H., and Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, 7:208–256.

Voss, J. L., Schendan, H. E., and Paller, K. A. (2010). Finding meaning in novel geometric shapes influences electrophysiological correlates of repetition and dissociates perceptual and conceptual priming. *NeuroImage*, 49:2879 – 2889.

Yonelinas, A. P. (1994). Receiver operating characteristics in recognition memory: Evidence for a dual-process model. *J. Exp. Psychol.: Learn. Mem. Cogn.*, 20:1341–1354.

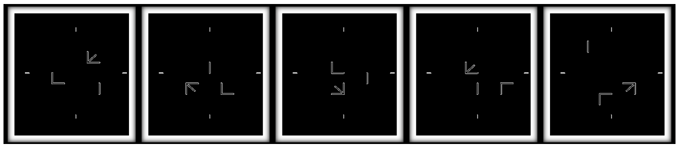

Figure 1: Five sample images.

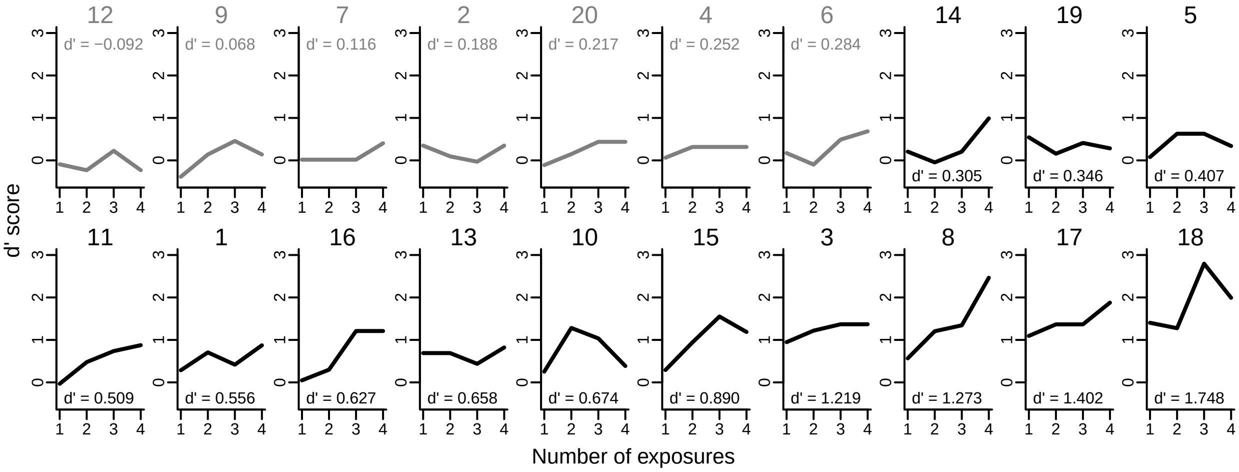

Figure 2: Estimated *d′* scores as a function of number of exposures *E* for underperforming (gray) and performing subjects (black).

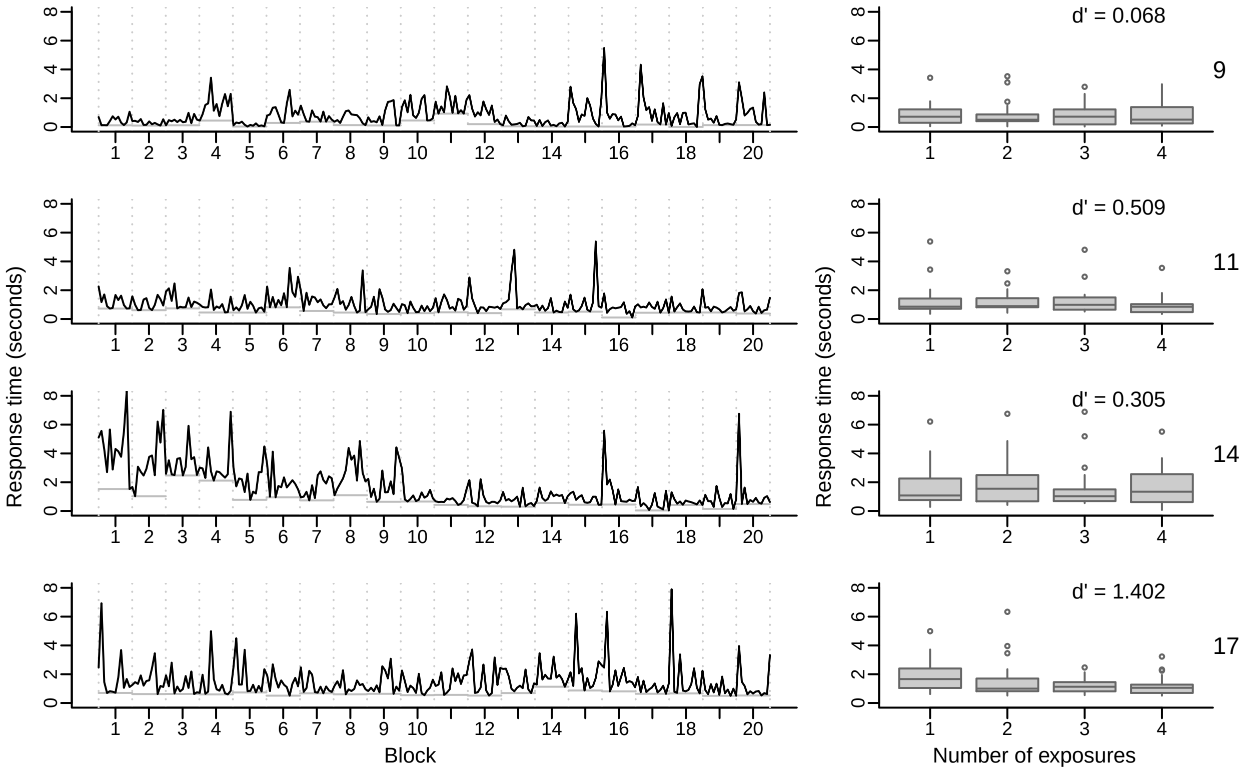

Figure 3: Left panels: RTs for subjects 9, 11, 14, and 17 over 20 blocks of 12 trials. The horizontal gray lines show the minimum RTs in each block and the vertical dotted lines delineate the blocks. Right panels: Boxplots of RT vs. exposures for the same subjects.

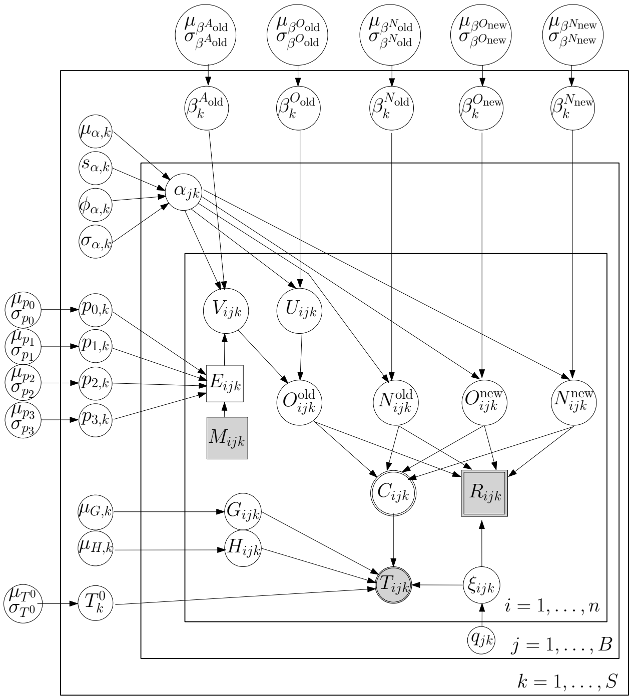
$\mu_{\beta^{A_{\text{old}}}}$ $\sigma_{\beta^{A_{\text{old}}}}$
$\mu_{\beta^{O_{\text{old}}}}$ $\sigma_{\beta^{O_{\text{old}}}}$
$\mu_{\beta^{N_{\text{old}}}}$ $\sigma_{\beta^{N_{\text{old}}}}$
$\mu_{\beta^{O_{\text{new}}}}$ $\sigma_{\beta^{O_{\text{new}}}}$
$\mu_{\beta^{N_{\text{new}}}}$ $\sigma_{\beta^{N_{\text{new}}}}$
$\beta_k^{A_{\text{old}}}$
$\beta_k^{O_{\text{old}}}$
$\beta_k^{N_{\text{old}}}$
$\beta_k^{O_{\text{new}}}$
$\beta_k^{N_{\text{new}}}$
$\mu_{\alpha,k}$
$s_{\alpha,k}$
$\phi_{\alpha,k}$
$\sigma_{\alpha,k}$
α_{jk}
μ_{p_0} σ_{p_0}
μ_{p_1} σ_{p_1}
μ_{p_2} σ_{p_2}
μ_{p_3} σ_{p_3}
$p_{0,k}$
$p_{1,k}$
$p_{2,k}$
$p_{3,k}$
V_{ijk}
U_{ijk}
E_{ijk}
M_{ijk}
O_{ijk}^{old}
N_{ijk}^{old}
O_{ijk}^{new}
N_{ijk}^{new}
C_{ijk}
R_{ijk}
$\mu_{G,k}$
$\mu_{H,k}$
G_{ijk}
H_{ijk}
T_{ijk}
ξ_{ijk}
μ_{T^0} σ_{T^0}
T_k^0
q_{jk}
$i = 1, \ldots, n$
$j = 1, \ldots, B$
$k = 1, \ldots, S$

Figure 4: Graphical representation of the trace-race model. Shaded nodes denote experimental constants or observable data and unshaded nodes denote unobservable RVs. Quantities in circles (squares) are real-valued (integer-valued). The arrows indicate distributional or logical dependence. Nodes with double outlines are functions of their parent nodes.

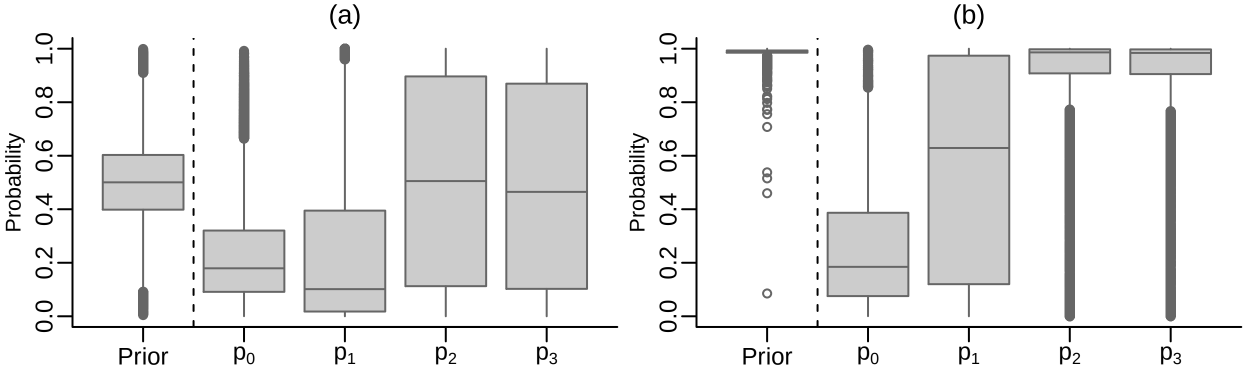

Figure 5: Boxplots of the posterior distributions for the p_ls under (a) a prior centered at 0.5 and (b) a very extreme prior concentrated near 1.

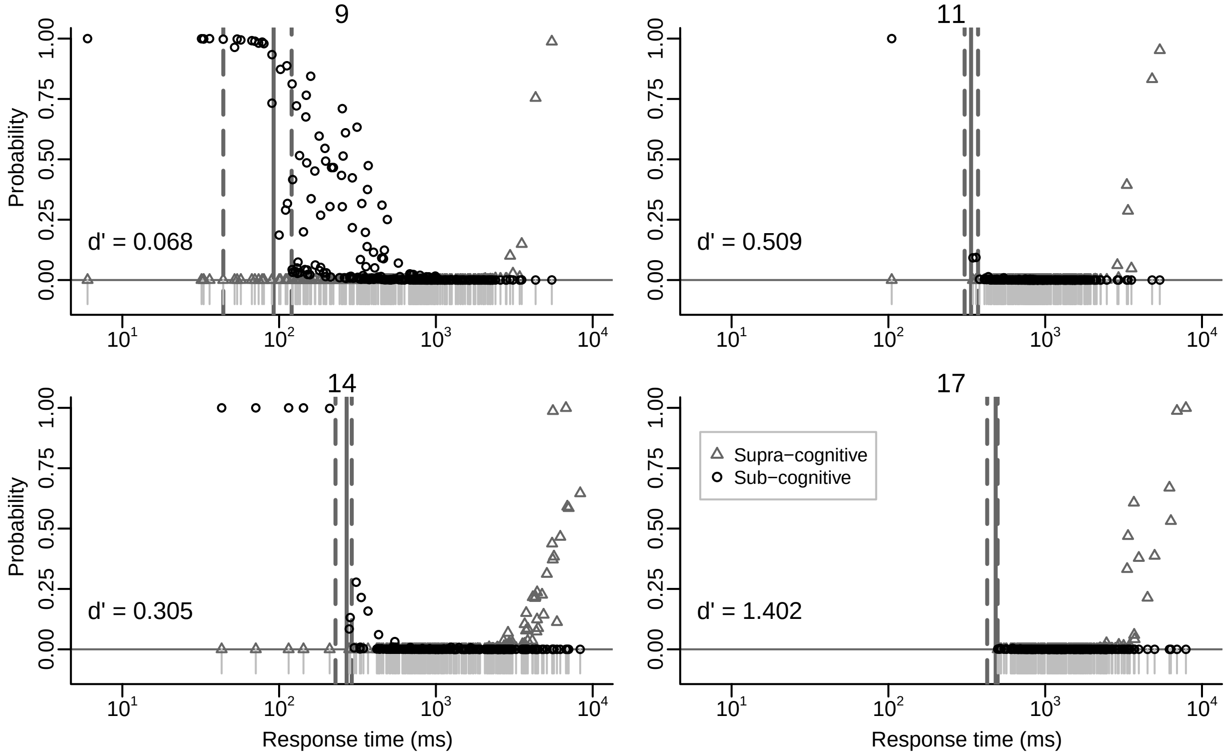

Figure 6: Each panel shows, for a subject k ($k = 9, 11, 14, 17$), the estimated posterior probability of being in the sub-cognitive state ($\xi_{ijk} = 0$; circles) and the supra-cognitive state ($\xi_{ijk} = 2$; triangles) versus the RT for each trial i and block j. The short gray vertical line segments plotted at the bottom of each panel correspond to the observed RTs. The vertical line denotes the posterior median of the residual time T_k^0 for each subject k and the dashed vertical lines denote a 95% posterior credible interval for T_k^0.

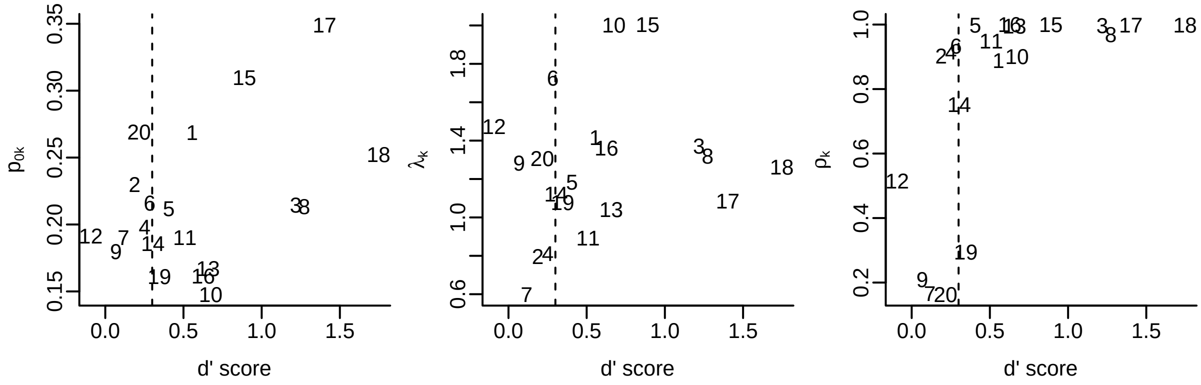

Figure 7: The posterior means of $p_{0,k}$, λ_k and ρ_k, versus the discriminability d' for the 20 subjects. The vertical dotted line is drawn at $d' = 0.3$, the threshold we used to divide performing from underperforming subjects.

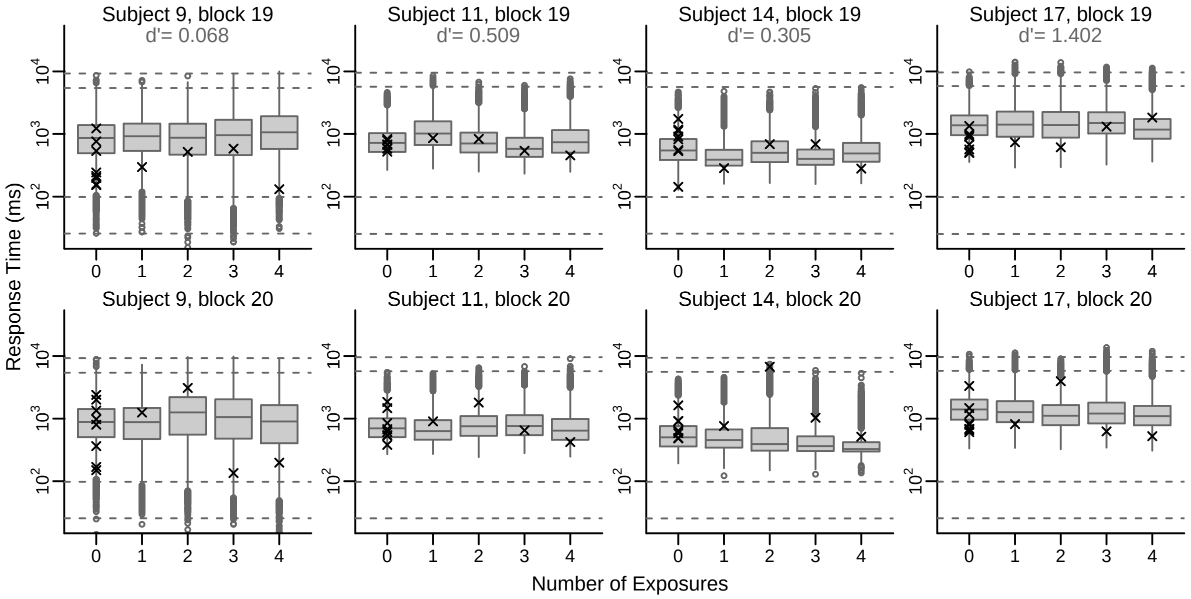

Figure 8: Log RTs versus number of exposures (black crosses), for four different subjects and blocks. The boxplots summarize the log predictive distribution of the cognitive processing times ($\log(C + T^0)$). The horizontal dashed lines above the boxplots indicate the 0.25 and 0.75 quantiles of the log predictive distribution for supra-cognitive processing times ($\log(H + T^0)$). The horizontal dashed lines below the boxplots indicate the 0.25 and 0.75 quantiles of the log predictive distribution for sub-cognitive processing times ($\log G$).

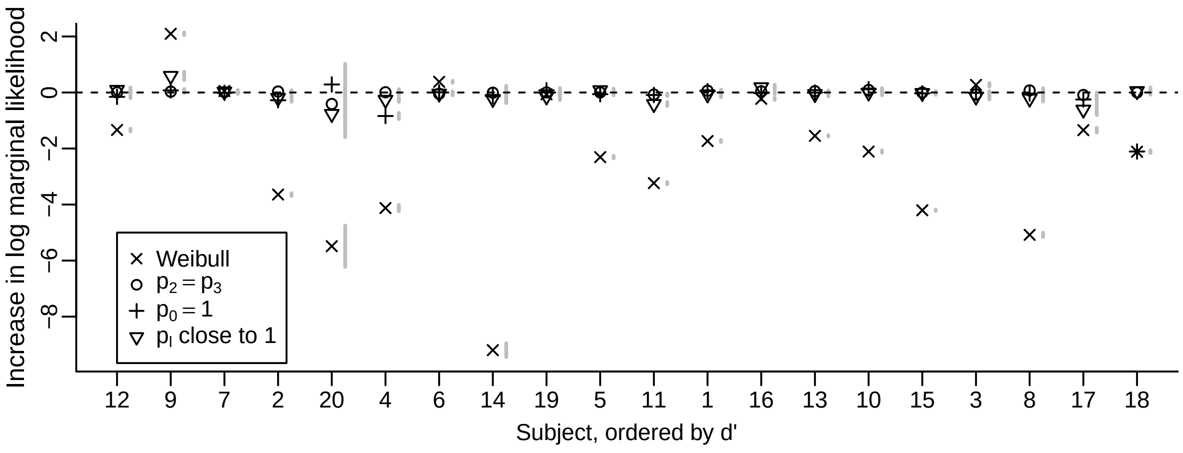

Figure 9: Increases in log-marginal likelihoods for the data from blocks 19 and 20 for each subject relative to the full minimum gamma (base) model. Differences larger than zero indicate a better fit for the alternative models (Weibull and restricted minimum gamma models) than the base model. The half-width of the error bars (vertical gray lines shown next to each symbol) are equal to 1.96 times the Monte Carlo estimates of standard error.