Modeling Memory Dynamics in Visual Expertise

Jeffrey Annis and Thomas J. Palmeri

Vanderbilt University

Correspondence should be addressed to Jeffrey Annis, 111 21st Ave S., 301 Wilson Hall, Nashville, TN 37240. E-mail: jeff.annis@vanderbilt.edu

Abstract

The development of visual expertise is accompanied by enhanced visual object recognition memory within an expert domain. We aimed to understand the relationship between expertise and memory by modeling cognitive mechanisms. Participants with a measured range of birding expertise were recruited and tested on memory for birds (expert domain) and cars (novice domain). Participants performed an old-new continuous recognition memory task whereby on each trial an image of a bird or car was presented that was either new or had been presented earlier with lag *j*. The Linear Ballistic Accumulator model (LBA Brown & Heathcote, 2008) was first used to decompose accuracy and response time into drift rate, response threshold, and non-decision time, with the measured level of visual expertise as a potential covariate on each model parameter. An expertise x category interaction was observed on drift rates such that expertise was positively correlated with memory performance recognizing bird images but not car images as old versus new. To then model the underlying processes responsible for variation in drift rate with expertise, we used a model of drift rates building on the Exemplar-Based Random Walk model (Nosofsky, Cox, Cao, & Shiffrin, 2014; Nosofsky & Palmeri, 1997), which revealed that expertise was associated with increases in memory strength and increases in the distinctiveness of stored exemplars. Taken together, we provide insight using formal cognitive modeling into how improvements in recognition memory with expertise are driven by enhancements in the representations of objects in an expert domain.

Experts are found in a wide variety of domains, such as chess (Chase & Simon, 1973), music (Wong & Gauthier, 2010), sports (Baker, Côté, & Abernehty, 2003), and physics (Chi, Feltovich, & Glaser, 1981). Our focus is visual experts, particularly those who have a marked ability to identify and categorize images of objects within their domain of expertise (Gauthier, Tarr, & Bub, 2010; Palmeri, Wong, & Gauthier, 2004; Shen, Mack, & Palmeri, 2014), such as dermatologists who categorize skin lesions as normal or cancerous, mycologists who categorize similar mushrooms as poisonous or edible, or birders who categorize hundreds of different species of birds. A key manifestation of visual expertise that we explore in the present work is its facilitating effect on visual recognition memory for images within an expert domain. Visual expertise is accompanied by increased visual short-term memory performance (Curby & Gauthier, 2007; Curby, Glazek, & Gauthier, 2009; Lorenc, Pratte, Angeloni, & Tong, 2014); for example, Curby et al. (2009) found that car expertise was significantly correlated with visual short-term memory capacity for cars but not faces. Visual expertise is also accompanied by increased visual long-term memory performance (Evans et al., 2011; Herzmann & Curran, 2011); for example, Evans et al. (2011) found that medical expertise led to significantly better long-term visual recognition memory for images within their medical expert domain compared to a novice domain.

The present work examines effects of visual expertise on both short-term and long-term recognition memory simultaneously. In short-term recognition memory tasks, a short array or sequence of study images are presented and then memory is tested soon after. In long-term recognition memory tasks, a longer array of study images is used and memory is tested after some delay. We combine these two types of memory tasks in an old-new *continuous recognition task* (e.g., Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993; Shepard & Teghtsoonian,

1961). On each trial, an image is presented and the participant judges that image as *old* or *new*, with old items having appeared previously with some lag *j* before the current trial. The inclusion of both short and long lags allows us to study both short- and long-term memory performance in the same task. To our knowledge, expert visual recognition memory has not been studied using a continuous recognition memory paradigm. The inclusion of lag within a memory task also provides additional constraints on our primary goal of modeling memory as a function of expertise.

We adopted a two-step cognitive modeling approach to understanding mechanistically how memory varies with visual expertise. This stepwise approach allowed us to first *measure how* underlying memory processes vary with visual expertise, and then *test why* these variations occur with expertise. Although past work has suggested that visual expertise might be driven by changes to memory representations and processes for images within a domain of expertise (Bukach, Gauthier, & Tarr, 2006; Palmeri et al., 2004), no formal cognitive model has been used to relate visual expertise and visual recognition memory.

We first applied a variant of the well-known class of *sequential sampling models* (Ratcliff & Smith, 2004). These models assume that evidence accumulates over time until a decision threshold is reached, at which point the response associated with that accumulator threshold is made. Variability in accumulation across trials allows these models to account naturally for both correct and error responses and the distributions of response times associated with those responses. Systematic modulation of the decision threshold, reflecting varying degrees of response threshold, allow these models to account naturally for speed-accuracy tradeoffs (e.g., Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998). While such sequential sampling models are general models of decision making and can be applied to a wide range of

perceptual and cognitive tasks, one of the best known early applications of these models was to memory (Ratcliff, 1978). By fitting these models to observed recognition memory performance, we can measure how model parameters associated with evidence, threshold, and non-decision time (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009) during short- and long-term memory decisions vary with visual expertise.

The particular sequential sampling model we chose to use is the *Linear Ballistic Accumulator* model (LBA; Brown & Heathcote, 2008). Like other models in its class, LBA assumes that once a stimulus has been perceptually encoded, evidence accumulates over time towards decision thresholds associated with alternative responses, which in the case of recognition memory are *old* and *new*. The LBA assumes a simple linear accumulation to threshold, with no within-trial variability in accumulation, but allows for between-trial variability in accumulation drift rate and starting point of accumulation; this simplification significantly speeds simulation of the model, which is especially important for the computationally-intensive Bayesian approaches we outline later (Annis & Palmeri, 2018).

We then asked why these parameters varied with expertise. Armed with a finding from the first modeling step that the rate of evidence accumulation – drift rate – driving memory decisions varies with expertise, we then tested sequential sampling models embodying alternative theories of drift rate in recognition memory that built on the *exemplar-based random walk model* (Nosofsky & Palmeri, 1997, 2015; Palmeri, 1997). These EBRW-based model variants (Nosofsky, Cao, Cox, & Shiffrin, 2014; Nosofsky, Cox, et al., 2014) make explicit alternative hypotheses about memory representation and processing assumptions and how these might vary with expertise. We chose EBRW among other theories of drift rates (Ashby, 2000; Logan, 2002; Smith & Ratcliff, 2009) because of prior work that explicitly relates EBRW and

LBA (Nosofsky, Cao, et al., 2014) and because EBRW has been shown to be able to account for memory performance in Sternberg short-term memory tasks (Nosofsky, Cox, et al., 2014) and long-term recognition memory tasks (e.g., Nosofsky & Palmeri, 2015; Nosofsky & Stanton, 2006). EBRW belongs to a class of similarity-based models. In the case of memory, old-new continuous recognition memory performance can be hypothesized to be a joint function of a test item's similarity to stored exemplars in memory, the overall strength with which exemplars are stored, and the modulation of model memory strength due to decay as a function of lag. EBRW allows us to test alternative hypotheses regarding why drift rates vary with visual expertise, specifically asking whether similarity, memory strength, or rate of decay vary with visual expertise. This modeling framework allows us to test a rich set of alternative hypotheses regarding the relationship between visual expertise and visual recognition memory in a formal manner.

Experiment

Many different types of visual experts have been studied. Here we focus on bird experts. There are several reasons for this choice:

Birding has been one of the canonical domains for studying how visual expertise modulates categorization (e.g., Tanaka & Taylor, 1991), memory (e.g., Herzmann & Curran, 2011), functional brain imaging (e.g., Gauthier, Skudlarski, Gore, & Anderson, 2000), and electrophysiology (Tanaka & Curran, 2001). There are practical reasons why birding has been such a popular choice (Shen et al., 2014). There are millions of people who birdwatch in the US (La Rouche, 2006), making it far easier to recruit from a large and diverse population of bird

experts than from highly specialized professional domains of visual expertise. Birders regularly participate in citizen science efforts, such as formal and informal bird counts, and are often quite willing to participate in experiments aimed at understanding their expertise. By contrast, for certain domains of expertise, like radiology and dermatology, it can be practically difficult to recruit large numbers of professionals to participate in experiments, and for other domains, like latent fingerprint examination and baggage screening, it can be bureaucratically burdensome or even illegal for those professionals to participate (e.g., Wolfe, Brunelli, Rubinstein, & Horowitz, 2013). In addition, compared to more esoteric or tightly controlled domains of expertise, there are hundreds of thousands of bird images readily obtainable online for use in visual cognition experiments.

Until recently, doing expertise research was somewhat challenging because expert participants would need to be recruited locally using advertisements posted in the neighboring community (e.g., Herzmann & Curran, 2011) or through peer recommendations by other identified experts (e.g., Tanaka & Taylor, 1991), thereby limiting the number of experts who could potentially be recruited. The advent of online web-based experiments has made expertise research far easier. In the case of birding, we have identified online hundreds of birding organizations across North America, many of which have granted us permission to advertise our experiments through their email list, newsletter, Facebook group, or website. These birding societies attract birders with a wide range of experience and expertise, from individuals with an interest in birds but little expertise, to those who make a living – or could – from their birding expertise. Our web-based experiments also capitalize on the growing literature that demonstrates the validity of online studies (Crump, McDonnell, & Gureckis, 2013; Germine et al., 2012; Gosling, Vazire, Srivastava, & John, 2004; Reimers & Maylor, 2005; Reimers & Stewart, 2007).

Variability in timing across different keyboards, browsers, and monitors has been shown to be relatively small in comparison to participant variability in response time (Crump et al., 2013) and several classic studies using response times have been replicated online (e.g., Crump et al., 2013), including classic studies of visual expertise (e.g., Shen et al., 2014).

To study visual expertise, it is important to estimate the location of experts along the expertise continuum. Self-report measures of expertise have been used in some past studies, especially where the goal is simply to establish a group of experts to compare to a group of novices (e.g. Evans et al., 2011). However, self-report alone is often an inadequate measure of expertise (Ericsson, 2006; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012). Therefore, we used an objective, quantitative measure of visual expertise. We chose to use one widely-used measure derived from a *subordinate matching* task (Gauthier, Curran, Curby, & Collins, 2003; Gauthier et al., 2000; Hagen, Vuong, Scott, Curran, & Tanaka, 2014; McGugin & Gauthier, 2010), which has been shown to predict both behavioral and brain changes that accompany visual expertise (e.g., Gauthier et al., 2000). On each trial of subordinate matching, the participant is sequentially presented pairs of birds (or cars) that are either the same or different species (or model) and the participant must distinguish between same versus different pairs. The discriminability (*d'*) for expert images (birds) versus non-expert images (cars) is used as the measure of birding expertise.

In the Method section below, we first describe the details regarding participant recruitment and the subordinate matching task of expertise. We then describe the details of the continuous recognition memory task.

Method

Participants

Fifty-four participants with a wide range of birding experience and expertise were recruited. Given the online nature of our experiments, we invited participation from a larger group of prospective participants who had previously registered on our web server, had completed a demographic questionnaire, and may have participated in previous online experiments in our lab. These participants had initially received advertisements or emails that had been directed at North American birding organizations. Of those who chose to accept our invitation, 8 self-reported as “beginner,” 20 as “intermediate,” and 26 as “expert”. All were given an opportunity to enter drawings for a 1/25 chance to win a $100 Amazon gift card. Twenty-one participants were female and 33 were male. Participants were between 22 and 72 years of age (M = 44.85, SD = 14.1). Participants gave informed consent to participate by electronically signing an informed consent form.

Subordinate Matching Task

The subordinate matching task was identical to that used in McGugin and Gauthier (2010). The stimulus set was composed of greyscale bird (passerines) and car (sedans) images. There were 112 images per category. Each image was 250 x 250 pixels. Participants completed 4 blocks of the subordinate matching task. Two blocks contained images of birds and two blocks contained images of cars. The order of the blocks and the order of the trials were kept constant across participants.

Participants completed 4 practice trials containing bird images before the first block containing birds and 4 practice trials containing car images before the first block containing cars. On each trial, participants were presented with an image of a bird or a car for 1000 ms followed by a mask presented for 500 ms. Immediately following the mask, a new bird or car image was presented that was either the same as or different from the previously presented species of bird or model of car. The task of the participant was to press the "d" key if the two images contained different species or models and "k" if the two images contained the same species or model. No corrective feedback was provided.

Continuous Recognition Memory Task

Immediately following the subordinate matching task, participants were presented with the instructions for the continuous recognition task. One-hundred color bird images (passerines) and 100 color car images (sedans) were used as stimuli in the continuous recognition task. Each image was 250 x 250 pixels. The sedan images were selected from the pool of images used by Herzmann and Curran (2011). We selected passerine bird images from a large pool of bird images collected from the internet. All images were cropped and placed on a blank background.

Participants completed 4 blocks of a continuous recognition task. Each block contained either images of birds or images of cars on a blank background. The category of the initial block was counterbalanced across participants. Each successive block contained a different category than the previous block. Each block contained 50 new images of which 40 were repeated. Five of the remaining 10 images that were not repeated were used as filler items to ease the computational burden of list creation and 5 images were used as load items presented on the first

5 trials of the block. Each successive presentation of an image was intervened by either 1 (lag 2) or 15 (lag 16) intervening images; within each block, 20 images were repeated at lags of 2 and 20 images were repeated at lags of 16. The order of the lists was randomized and generated anew for each participant such that lag 2 and lag 16 images were distributed uniformly across the list. The same image set was used for each participant. In order to ensure that lag was not confounded with list position, the list was divided in half and the frequency of lag 2 and the frequency of lag 16 images was computed for each half of the list. A chi-squared test for uniformity was performed on the resulting frequency tables. Lists that failed the test at the .05 significance level were rejected and a new list was created in its place. We used lags of 2 and 16 because we had a limited amount of time that participants would be willing to complete an online experiment.

On each trial of continuous recognition, the task was to press the "d" key if the current image was previously studied (old) and to press the "k" key if this was the first time the image was presented (new). The task was self-paced. Participants were instructed to place one left finger on the "d" key and one right finger on the "k" key. Eight practice trials with feedback were presented before beginning the actual experiment. Following the practice trials, participants were presented with each block. No feedback was provided during actual experimental trials. Following each block, the overall accuracy for the most recent block was shown to the participant.

Results and Discussion

We trimmed the data such that responses greater than 6 s or less than 150 ms (~ 0.03% of responses) were omitted from the analysis. The difference in *d'* between the bird and car

categories in the Subordinate Matching task was used as a measure of expertise (Herzmann & Curran, 2011). We refer to this measure as the *expertise index* or $\Delta d'$ and use this as a covariate in all subsequent analyses. In order to control for any age effects, we also included age as a covariate. We found that, in our sample of participants, age was negatively correlated with expertise, $r(52) = -.30$, $p < .05$ and that RTs for hits and correct rejections increased with age, $r(52) = .41$, $p < .05$ and $r(52) = .37$, $p < .05$, respectively; however, we found that age did not interact with category (bird or car) for any dependent measure and, therefore, we do not explicitly report analyses regarding age, but still include it as a covariate in our analyses. A detailed statistical analysis can be found in the Appendix. We report the main findings below.

The expertise index, $\Delta d'$, ranged from -1.19 to 3.42 ($M = 1.53$, $SD = .93$); two participants had $\Delta d'$ scores less than or equal to 0, indicating greater car than bird expertise, with the remaining having $\Delta d'$ scores greater than 0. A 2 (category: bird vs. car) x 2 (lag: 2 vs. 16) repeated-measures ANCOVA was conducted with the expertise index ($\Delta d'$) and participant age as covariates. The left panel of Figure 1 shows accuracy as measured by *d'* was greater in the bird condition ($M = 2.28$, $SD = .80$) than in the car condition ($M = 1.34$, $SD = .40$), $F(1,51) = 107.04$, $p < .0001$, and was greater for lags of 2 ($M = 1.82$, $SD = .51$) than for lags of 16 ($M = 1.64$, $SD = .49$), $F(1,51) = 12.69$, $p < .001$. The category x lag interaction was not significant, $F(1,51) = 3.44$, $p = .069$. The right panel of Figure 1 shows *d*' plotted as a function of expertise and lag for each category. There was a significant main effect of expertise, $F(1,51) = 4.18$, $p < .05$, and a significant expertise x category interaction, $F(1,51) = 6.92$, $p < .05$. Simple linear regression revealed that expertise predicted *d'* for bird images ($\beta = .33$, $p < .01$, adjusted $R^2 = 0.13$), but not for car images, ($\beta = .01$, $p = .926$, adjusted $R^2 = 0.00$). Consistent with Herzmann & Curran (2011), there was a facilitating effect of expertise on recognition accuracy as measured

by *d'* for images within the expert-domain only. This facilitating effect was not observed to significantly vary as a function of lag.

In summary, visual expertise for birds had a facilitating effect on continuous recognition performance for bird images, but not for car images. Specifically, we observed increased accuracy as measured by *d'* for expert domain stimuli. We also conducted an analysis on response time (see appendix), but did not observe such an interaction in the response times. Although we did not observe covariation in response time with expertise in memory, they nevertheless add important constraints for the model we develop in the next section in which we jointly model accuracy and response time in the LBA framework to investigate the relationship between visual expertise and continuous recognition memory performance.

Modeling Methods

In this section, we measure how key psychological mechanisms vary with visual expertise using the LBA (Brown & Heathcote, 2008). An illustration of the LBA is shown in Figure 2. The LBA assumes that evidence for each response type, *old* and *new*, is accumulated over time in a linear fashion. A decision is made when the amount of evidence accumulated reaches a pre-determined threshold b. The rate at which evidence accumulates for each response type is given by the drift rates, d_{old} and d_{new}. The drift rates are assumed to be drawn from normal distributions with mean v_{old} or v_{new} and standard deviation s. The difference in v_{old} given old vs. new stimuli we refer to as v', where $v' = v_{old|old} - v_{old|new}$. Increases in v' reflect increases in the ability to discriminate between old and new images, conceptually akin to *d'* from signal detection theory.

The LBA also assumes that the starting point of evidence accumulation, *a*, varies between trials and is drawn from a uniform distribution between 0 and *A*. Between-trial variability allows the model to capture important differences between the correct and error response time distributions. The difference, *k*, between *A* and *b,* we refer to as the *relative threshold*. Differences in *k* reflect differences in response threshold, which is assumed to be at least partially under the control of the participant. As the participant decreases their response threshold, this increases the likelihood that an incorrect response will be made. Lastly, the time it takes to perceptually encode the stimulus and execute the motor response is given by the non-decision-time parameter[1], τ.

The relationship between visual expertise and three key components of the LBA model were examined: drift rate (*v'*), relative threshold (*k*), and non-decision time (τ). If visual expertise is associated with increases in the quality of evidence upon which recognition memory decisions are made, then *v'* should increase with expertise in the bird condition compared to the car condition. If visual expertise is associated with increases in the efficiency of perceptual processing, then τ should decrease with expertise. If visual expertise is associated with differences in response threshold, then *k* should differ with expertise.

We chose to implement our modeling in a Bayesian hierarchical framework (e.g., see Annis, Miller, & Palmeri, 2017; Annis & Palmeri, 2018; Kruschke, 2014; Lee & Wagenmakers, 2014). Bayesian hierarchical models have been shown to increase the stability of the parameter estimates when there are low numbers of observed data points per participant and relatively high numbers of participants (Katahira, 2016; Kruschke & Vanpaemel, 2015). This is an important advantage given our online web-based experiments. One challenge of conducting web-based experiments is balancing the length of an experiment with the potential attrition rate. In an

uncontrolled web-based environment, it is not unlikely for a participant to simply quit an experiment with a click of a button, making them unlikely to participate in long or tedious experiments.

This challenge can prove especially problematic when it comes to fitting models like LBA, which under conventional modeling methods require hundreds of trials per condition per participant to yield sufficient distributions of error and correct response needed to fit these models. In a sense, our modeling is a proof of concept that RT models like LBA can be fitted in a Bayesian hierarchical framework to data in online experiments with limited numbers of observations per participant and in our case that this modeling can reveal something about the mechanisms underlying memory in visual domain experts.

All models were implemented using the Stan probabilistic programming language (Carpenter et al., 2016). We initially attempted to use the default MCMC algorithm in Stan, called NUTS, but found it required prohibitively long sampling times; we believe this was due to the high complexity of the LBA likelihood function and model structure. Therefore, we decided to use an alternative algorithm in Stan called *automatic differentiation variational inference* (ADVI; Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2017), which was developed in order to scale Bayesian inference to big data and complex models. Variational inference minimizes the Kullback-Liebler divergence between the actual posterior and an approximation of the posterior by maximizing the *evidence lower bound* of the model (the expected joint density minus the entropy under the approximation) via stochastic gradient ascent. ADVI stops when the stochastic gradient ascent procedure can no longer improve the evidence lower bound according to a pre-determined tolerance. Samples from the approximate posterior can then be drawn. Posterior estimates obtained with ADVI have been shown to accurately reflect those obtained with NUTS

(Kucukelbir et al., 2017). For our model fits, we used the fullrank-ADVI algorithm and drew 1000 samples following the completion of stochastic gradient descent. We set the relative tolerance to .003 (a value we found through pilot work that led to convergence of stable estimates) and used default parameters settings otherwise.

Subordinate Matching Model

First, the subordinate matching task was modeled using Bayesian Signal Detection Theory (Green & Swets, 1966; Lee, 2008b). For each participant *i* in condition *j*, the total number of hits (correct "same" responses) and false alarms (incorrect "same" responses) are assumed to follow a binomial distribution:

$$H_{ij} \sim \text{Binomial}(h_{ij}, T),$$

$$F_{ij} \sim \text{Binomial}(f_{ij}, L),$$

where h_{ij} is the hit rate, f_{ij} is the false alarm rate, and T is the number of targets (same trials) and L is the number of lures (different trials). The hit and false alarm rates are parameterized in terms of sensitivity, d_{ij}, and bias, c_{ij}:

$$h_{ij} = \Phi\left(\frac{1}{2}d_{ij} - c_{ij}\right),$$

$$f_{ij} = \Phi\left(-\frac{1}{2}d_{ij} - c_{ij}\right),$$

where Φ is the CDF of the standard normal distribution. Priors are placed on d_{ij} and c_{ij} (following Lee, 2008a) such that

$$d_{ij} \sim \text{Normal}(\mu_j^d, \sigma_j^d),$$

$$c_{ij} \sim \text{Normal}\left(\mu_j^c, \sigma_j^c\right),$$

$$\mu^d, \mu^c \sim \text{Normal}(0,2),$$

$$\sigma^d, \sigma^c \sim \text{Gamma}(1,1).$$

The difference in sensitivity between the car category and bird category for each participant, Δd_i, is used as an index of expertise (Herzmann & Curran, 2011).

$$\Delta d_i = d_i^{bird} - d_i^{car},$$

where d_i^{car} corresponds to the car category and d_i^{bird} corresponds to the bird category. We show below how Δd_i is used as a potential covariate within the LBA.

The subordinate matching model and the LBA were fit simultaneously to the entirely of the observed data. Thus, the memory task informed parameter estimations in the subordinate matching task and vice versa. This is one advantage of using Bayesian hierarchical modeling methods.

Linear Ballistic Accumulator

For each participant *i* in category *j* given lag *l*, response time and response choice pairs, $\boldsymbol{RT}_{ijl}$, are distributed according to the LBA:

$$\boldsymbol{RT}_{ijl} \sim \text{LBA}(A, s_i, v'_{ijl}, k_{ij}, \tau_{ij}),$$

where *A* is fixed to 1 in order to make the model identifiable (Donkin, Brown, & Heathcote, 2011) and s_i is the drift rate variability with the following priors:

$$s_i \sim \text{Normal}(\mu^s, \sigma^s),$$

$$\mu^s, \sigma^s \sim \text{Normal}(1,1).$$

$$\sigma^s \sim \text{Gamma}(1,1).$$

All priors are roughly based on those that have been used in previous modeling work with the LBA (e.g., Turner, Sederberg, Brown, & Steyvers, 2013). We model v'_{ijl} directly and make $v_{ijl}^{Old|Old}$ a deterministic parameter:

$$v_{ijl}^{Old|Old} = v'_{ijl} + v_{ijl}^{Old|New},$$

where

$$v_{ijl}^{Old|New} \sim \text{Normal}\left(\mu_{jl}^{v^{Old|New}}, \sigma_{jl}^{v^{Old|New}}\right),$$

$$\mu_{jl}^{v^{Old|New}} \sim \text{Normal}(2,2) \in (0,\infty),$$

$$\sigma_{jl}^{v^{Old|New}} \sim \text{Gamma}(1,1).$$

[note: the notation, $\in [0,\infty)$, represents a truncated normal between 0 and infinity]. The remaining parameters are regressed on Δd_i (index of expertise for participant *i*):

$$v'_{ijl} \sim \text{Normal}\left(\mu_{jl}^{v'} + \beta_{jl}^{v'}\Delta d_i, \sigma_{jl}^{v'}\right),$$

$$k_{ij} \sim \text{Normal}\left(\mu_j^k + \beta_j^k \Delta d_i, \sigma_j^k\right) \in (0,\infty),$$

$$\tau_{ij} \sim \text{Normal}\left(\mu_j^\tau + \beta_j^\tau \Delta d_i, \sigma_j^\tau\right) \in (0,\infty),$$

where μ is the grand mean, σ is the standard deviation, and β is the regression coefficient. Priors on the grand means and standard deviations were mildly informative (Turner et al., 2013):

$$\mu_{jl}^{v'} \sim \text{Normal}(0,1)$$

$$\mu_j^k \sim \text{Normal}(1,2) \in [0,\infty)$$

$$\mu_j^\tau \sim \text{Normal}(.5,1) \in [0,\infty)$$

$$\sigma_{jl}^{v'}, \sigma_j^k, \sigma_j^\tau \sim \text{Gamma}(1,1).$$

The difference in the regression coefficients between category conditions is modeled directly as $\Delta\beta$ and the regression coefficient for the bird image condition becomes a deterministic parameter. For each regression weight, we have:

$$\beta_{bird,l}^{v'} = \Delta\beta_{l}^{v'} + \beta_{car,l}^{v'},$$

$$\beta_{bird}^{k} = \Delta\beta^{k} + \beta_{car}^{k},$$

$$\beta_{bird}^{\tau} = \Delta\beta^{\tau} + \beta_{car}^{\tau}.$$

A Savage-Dickey ratio test (Dickey, 1971; Wetzels, Grasman, & Wagenmakers, 2010) was performed on each $\Delta\beta$ in order to derive the Bayes factor for the expertise x category interaction. Priors on the regression coefficients follow standard normal distributions (Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

$$\Delta\beta_{l}^{\Delta v}, \Delta\beta^{k}, \Delta\beta^{\tau}, \beta_{car,l}^{v'}, \beta_{car}^{k}, \beta_{car}^{\tau} \sim \text{Normal}(0,1).$$

Modeling Results

Before analyzing the relationship between visual expertise and the model parameters, we first determined whether the model was able to provide a reasonable account of the data. Our primary goal was to determine whether parameters varied with visual expertise, not to fit the data perfectly. Although we had quite low numbers of observations per participant compared to traditional fits of RT models, we found that the LBA model accounted for most of the data quite well with most correlations between predicted and observed well above .90. The model only had trouble on a subset of the data, the missed targets, which was probably because of the relatively

low number of observed misses (correlations between predicted and observed ranged from .56 to .81).

Given that the model adequately accounted for the data, we then tested whether certain parameters varied with visual expertise. Specifically, we tested whether expertise interacted with category for each parameter using the Bayesian framework. The null hypothesis, $\mathcal{H}_0$, states that $\beta_{car} - \beta_{bird} = 0$, and the alternative hypothesis, $\mathcal{H}_1$, states that $\beta_{car} - \beta_{bird} \neq 0$. The Bayes factor indicates how much more likely the data are under the null hypothesis than the alternative hypothesis. When the Bayes factor is greater than 3 (Kass & Raftery, 1995), this is conventionally interpreted as positive support for the null hypothesis (in our case, no effect of expertise), and when the Bayes factor is less than 1/3, this indicates positive support for the alternative (in our case, an effect of expertise). We present a detailed discussion of the model fits followed by inferences on parameters.

Model Fits

Panel A of Figure 3 shows that the model was able to capture the increased hit rates in memory for bird images (expert) compared to the car images (novice). The model was also able to capture the steeper decline in hit rates as a function of lag for cars compared to birds. Panel B of Figure 3 shows the individual-level observed and predicted hit rates as a function of image category and lag. For recognition of birds, most participants had a high degree of accuracy and are clustered in the top right corners for lags of 2 and 16, which the model was able to capture (r = .96 and r = .98, respectively). Panel C of Figure 3 shows the model successfully captured the decrease in false alarm rates for birds compared to the cars at the group-level. Panel D of Figure

3 shows observed and predicted false-alarm rates for each category at the individual-level. The model was able to capture the overall increase in false alarm rates for cars (r = .96) while also accounting for more variation across participants for birds (r = .98).

Panel A of Figure 4 shows the model was able to successfully capture the group-level response time quantiles for hits. Quantiles were computed by taking the response time for which 10%, 50%, and 90% of the response times fell below. Note that the group-level data for misses, especially for birds, were quite noisy due to the relatively low number of observed misses, and the model fits slightly suffer because of this. Panel B of Figure 4, shows the 10%, 50%, and 90% individual-level predicted and observed response time quantiles for old items plotted as a function of category, response type, and lag. As was the case for the group-level response time data for misses, the low number of missed trials limits the ability of the model to produce perfect predictions. This is especially the case for birds, where the number of misses is very low (r = .61 for lags of 2 and r = .56 for lags of 16). For cars, this is less of an issue because of the lower accuracy and hence higher number of misses (r = .81 for lags of 2 and r = .56 for lags of 16). However, for both the bird and car category, the model accurately predicts response times for hits for lags of 2 (r = .97 and r = .96, respectively) and 16 (r = .94 and r = .92, respectively). Panel C of Figure 4 shows the observed and predicted response time quantiles for new items for each category and response type. The model captures the overall response times for correct rejections and false alarms reasonably well. Panel D of Figure 4 shows the model was able to accurately predict participant-level response time quantiles of correct rejections and false alarms for both birds (r = .97 and r = .88, respectively) and cars (r = .96 and r = .92, respectively).

Overall, the model predictions were qualitatively satisfactory and captured all the major trends at both the group level and individual level. Quantitatively, the correlations between

predicted and observed hits and false alarm rates were very high, > .95. Correlations between predicted and observed response times were also high, >.87, except for response times for misses, where we observed moderate correlations, > .55.

What LBA Model Parameters Covary with Expertise?

Having adequately accounted for continuous recognition memory performance overall, we now move on to our main goal, to determine the relationship between visual expertise and model parameters. Figure 5 shows each LBA model parameter's predicted value as a function of the expertise index ($\Delta d'$) and category (bird vs. car). Predictions were generated by obtaining the mean of 500 samples drawn from the distributions in Eq. 1, 2, and 3 for each posterior sample. We then obtained the grand mean over these means (solid line) and the 95% highest density interval (HDI; dotted line). This procedure was done over a fine-grained sequence of $\Delta d'$ values. For each parameter, we tested whether expertise differentially covaried with category. Bayes factors greater than 3 indicate positive support for the null hypothesis of no modulation of a model parameter with expertise for bird images (i.e., a null expertise x category interaction). Bayes factors less than 1/3 indicate positive support for the alternative hypothesis of significant modulation of a model parameter with expertise for bird images (i.e., a significant expertise x category interaction). Bayes factors falling between 1/3 and 3 are conventionally interpreted as not indicating positive support for either hypothesis. Bayes factors are denoted as *BF* throughout.

We acknowledge that Bayes factors rely on the careful specification of priors that take into account, for example, the expected scale of the parameters. Given that our extension and application of the LBA is fairly novel, the grounds by which we specified priors were limited.

Therefore, in addition to reporting the Bayes factors, we also report the 95% HDI for each of the $\Delta\beta$ parameters the Bayes factor is based on, which is much less sensitive to the priors. We found agreement between the Bayes factors and HDIs on parameter estimates: When the Bayes factor indicated a null effect, the 95% HDI for $\Delta\beta$ included zero, and when the Bayes factor indicated a non-null effect, the HDI did not include zero.

The top panel shows the predicted response threshold, *k*, as a function of the expertise index, $\Delta d'$, for both categories. The data provided strong evidence for a null expertise x category interaction, BF = 21.29, $\Delta\beta^k$ HDI [-.06, .09]; the posterior distributions of the regression coefficients on the expertise index for the car category, β^k_{car}, and bird category, β^k_{bird}, both had 95% HDIs close to zero, [-.12, .04] and [-.11, -.01], respectively. Non-decision time, τ, also did not show an interaction between category and expertise, BF = 115.64, $\Delta\beta^\tau$ HDI [-.02, .02]; the posterior distributions of the regression coefficients on the expertise index for the car category, β^τ_{car}, and bird category, β^τ_{bird}, had 95% HDIs centered around or close to zero, [-.03, -.005] and [-.01, .03], respectively. Neither response threshold nor non-decision time varied with expertise for expert-domain images (birds).

v_2' is plotted as a function of expertise and category; recall that $v_2{}'$ is the difference in accumulation rates towards the old response between old stimuli at lags of 2 and new stimuli. Thus, $v_2{}'$ can be conceptualized as a discriminability measure much like *d'*. The data provided strong evidence for a category x expertise interaction, BF = 1/333, $\Delta\beta_2^{v'}$ HDI [.14, .32]. The posterior distributions of the regression coefficients on the expertise index in the car category, $\beta^{v_2'}_{car}$, and bird category, $\beta^{v_2'}_{bird}$, had 95% HDIs that fell above zero, [.02, .11] and [.21, .41], respectively. Thus, v_2' increased with expertise more so for bird images than for car images. v_{16}'

also showed an interaction between category and expertise, BF = 1/10718, $\Delta\beta_{16}^{v'}$ HDI [.20, .35]. The posterior distribution for the regression coefficient on the expertise index for the car category, $\beta_{car}^{v'_{16}}$, had a 95% HDI that included zero [-.01, .06], while the coefficient for the bird category, $\beta_{bird}^{v'_{16}}$, had a 95% HDI that fell below zero, [.19, .37]. Thus, v'_{16} increased with increases in expertise in the bird category, but the data provide little to no evidence for this in the car category. Increases in visual expertise were accompanied by changes in v'_2 and v'_{16} for images in the domain of expertise, indicating an increase in the quality of evidence entering into the decision process with an increase in visual expertise.

In addition, response threshold was shown to decrease with increases in visual expertise for both categories with no interaction. Recall from the statistical analyses, visual expertise was found to be negatively correlated with age, and age was found to be positively correlated with increased RTs. Therefore, the increases in response threshold with decreases in expertise might be due to age-related slowing. However, we merely speculate that this is the case because we did not include age as an explicit covariate in the model in order to reduce model complexity. More importantly, this result indicates that simple changes in threshold are not driving the increases in performance observed with increased visual expertise.

Taken together, our results suggest that greater drift rates in memory decisions may accompany greater visual expertise for images in an expert domain. It is also important to note that the absence of an effect of expertise on non-decision time (τ) does not imply an absence of an effect of expertise on perceptual encoding mechanisms and perceptual representations. To the contrary, changes in drift rate likely reflect such changes (e.g., Palmeri & Cottrell, 2009; Palmeri & Tarr, 2008; Palmeri et al., 2004) because the quality of visual memory representations that

drive recognition memory decisions depend on the quality of perceptual representations. To further investigate the underlying memory processes that may be driving these differences in drift rates with expertise, we tested alternative hypotheses regarding visual expertise by modeling drift rates in an exemplar-based framework.

Modeling Drift Rates

So far, we showed that drift rate (and not threshold or non-decision time) in the LBA for memory decisions about expert-domain images increases with increases in visual expertise. Here, we test alternative hypotheses for how a model-based decomposition of drift rate into theoretical subcomponents might vary with visual expertise to more deeply understand the nature of expertise-driven changes in memory mechanisms. We extend a model developed by Nosofsky, Cox, et al. (2014) based on the Exemplar-Based Random Walk model (EBRW; Nosofsky & Palmeri, 1997, 2015; Palmeri, 1997). Nosofsky, Cox, et al. developed this model to account for short-term and long-term recognition memory in a Sternberg task, making it straightforward to extend as model of continuous recognition memory.

The model assumes that on each trial of a continuous recognition memory task, a corresponding memory trace, an exemplar, is stored in memory. Illustrated in Figure 6 are three such exemplars currently stored in memory, e_1, e_2, and e_3; this makes the current trial 4. The model assumes that memory decisions are based on the activation produced from a match between the test cue, the item currently being judged as *old* or *new*, and the exemplars stored in memory. The activation value for each stored exemplar is depicted as ω_1, ω_2, and ω_3 in the figure. The familiarity of the test cue is a monotonically increasing joint function of the memory

strength (α) associated with each exemplar and the similarity (η) between the test cue and the activated set of exemplars. Memory strength is assumed to asymptotically decay (γ) as a function of lag. The summed activation of exemplars in turn drives an accumulation process in the LBA in which the drift rate corresponding to the old response, v_{old}, is proportional to this summed activation and v_{new} is constrained to be equal to 1-v_{old} (Nosofsky, Cox, et al., 2014). Of course, LBA is not a random walk model, and we are not using the random walk component of the EBRW. We are simply taking the front-end, the "theory of drift rates", from EBRW and marrying it with the LBA – creating, in a sense, an EB-LBA.

The model provides three key parameters to relate changes in visual expertise to changes in recognition memory performance: those associated with memory decay, overall memory strength, and similarity. If visual expertise is associated with changes in memory decay, then memory decay should decrease with expertise for the bird category and not the car category. If visual expertise is associated with increases in overall memory strength, then memory strength should increase with expertise. Lastly, if visual expertise is associated with changes in the distinctiveness of stored exemplars, then the similarity parameter should decrease with expertise.

Exemplar-Based Random Walk Model

Here we explicate details of the model outlined above. The model assumes that for participant *i* in category *j*, response time and choice pairs, $\boldsymbol{RT}_{ijl}$, are distributed according to the LBA using the drift rates, v_{ijl}^{Old}, defined by EBRW:

$$\boldsymbol{RT}_{ijl} \sim \text{LBA}(v_{ijl}^{Old}, 1 - v_{ijl}^{Old}, A, s_i, k_{ij}, \tau_{ij}),$$

Following Nosofsky, Cox, et al. (2014), drift rates are driven by the summed activation of a cue item to exemplars stored in memory:

$$v_{ijl}^{Old} = \frac{\Omega_{ijl}}{\Omega_{ijl} + r_i},$$

$$v_{ijl}^{New} = 1 - v_{ijl}^{Old},$$

where r_i is the activation of background elements used as a criterion to compare the activation produced by the match between the cue and the exemplars stored in memory, and Ω_{ijl} is the total activation of all exemplars entered into the memory match process. We note that the connection between the rate of accumulation in the LBA and rate equations derived for EBRW (Nosofsky & Palmeri, 1997) is one by analogy only (Nosofsky, Cao, et al., 2014) and do not make a claim regarding the formal mathematical relationship between the two. The total activation is assumed to include the exemplars from the previous test position to the maximum lag in the design, *G* (the lag between the current position and the first position in the list):

$$\Omega_{ijl} = \sum_{g=1}^{G} \omega_{ijg},$$

where ω_{ijg} is the activation of a stored exemplar at lag *g*, e_g. This activation is assumed to be the result of a matching process between the current test cue, *t*, and the stored exemplar e_g. The activation is governed by the memory strength m_{ijg} of the stored exemplar scaled by the similarity $\Phi(\eta_{ij})$ between the current test cue and the exemplar. When the current test cue is the same as the stored exemplar e_g, similarity is set to 1. When the current test cue is not the same as the stored exemplar, then similarity is modeled as a real number, η, transformed by the CDF of

standard normal distribution[2], Φ, such that $0 < \Phi(\eta) < 1$. Thus, the activation of a stored exemplar is given by:

$$\omega_{ijg} = \begin{cases} m_{ijg} & , \quad e_g = t \\ \Phi(\eta_{ij}) m_{ijg}, & e_g \neq t, \end{cases}$$

Memory strength, m_{ijg}, is assumed to be a decaying function of lag, g:

$$m_{ijg} = \alpha_{ij} + \lambda_{ij} g^{-\gamma_{ij}}.$$

where α_{ij} is the memory strength asymptote, λ_{ij} is a scaling parameter, and γ_{ij} models memory decay. We regress the following EBRW parameters on Δd_i:

$$\alpha_{ij} \sim \text{Normal}\left(\mu_j^{\alpha} + \beta_j^{\alpha} \Delta d_i, \sigma_j^{\alpha}\right),$$

$$\eta_{ij} \sim \text{Normal}\left(\mu_j^{\eta} + \beta_j^{\eta} \Delta d_i, \sigma_j^{\eta}\right),$$

$$\gamma_{ij} \sim \text{Normal}\left(\mu_j^{\gamma} + \beta_j^{\gamma} \Delta d_i, \sigma_j^{\gamma}\right).$$

We used priors on the means based on best-fitting parameter values from Nosofsky, Cox, et al. (2014):

$$\mu_j^{\gamma}, \mu_j^{\alpha} \sim \text{Normal}(1,2) \in (0, \infty),$$

$$\mu_j^{\eta} \sim \text{Normal}(-1.5, 0.5).$$

Priors on standard deviations were weakly informative:

$$\sigma_j^{\gamma}, \sigma_j^{\alpha}, \sigma_j^{\eta} \sim \text{Gamma}(1, .5).$$

We model the difference between the regression coefficients in the bird and car conditions as:

$$\beta_{bird}^{\gamma} = \Delta\beta^{\gamma} + \beta_{car}^{\gamma},$$

$$\beta_{bird}^{\alpha} = \Delta\beta^{\alpha} + \beta_{car}^{\alpha},$$

$$\beta_{bird}^{\eta} = \Delta\beta^{\eta} + \beta_{car}^{\eta}$$

Priors on the regression coefficients follow standard normal distributions (Rouder et al., 2017).

$$\Delta\beta^{\gamma}, \Delta\beta^{\alpha}, \Delta\beta^{\eta}, \beta_{car}^{\gamma}, \beta_{car}^{\alpha}, \beta_{car}^{\eta} \sim \text{Normal}(0,1).$$

The remaining parameters of the EBRW not entered into the regression have the following priors based on values from Nosofsky, Cox, et al. (2014):

$$r_i \sim \text{Normal}(\mu^r, \sigma^r),$$

$$\mu^r \sim \text{Normal}(1,2),$$

$$\lambda_{ij} \sim \text{Normal}\left(\mu_j^{\lambda}, \sigma_j^{\lambda}\right),$$

$$\mu_j^{\lambda} \sim \text{Normal}(1,1).$$

Priors on standard deviations were weakly informative:

$$\sigma^r \sim \text{Gamma}(1,0.5),$$

$$\sigma_j^{\lambda} \sim \text{Gamma}(1,1).$$

For the LBA, we regress threshold and perceptual encoding parameters on the expertise score Δd_i:

$$k_{ij} \sim \text{Normal}\left(\mu_j^k + \beta_j^k \Delta d_i, \sigma_j^k\right) \in [0, \infty),$$

$$\tau_{ij} \sim \text{Normal}\left(\mu_j^{\tau} + \beta_j^{\tau} \Delta d_i, \sigma_j^{\tau}\right) \in [0, \infty),$$

Priors for the LBA were chosen based on previous hierarchical modeling work with the LBA (e.g., Turner et al., 2013) and are the same as those we used in the previous LBA model:

$$\mu_j^k \sim \text{Normal}(1,2) \in (0, \infty)$$

$$\mu_j^{\tau} \sim \text{Normal}(.5,1) \in (0, \infty)$$

$$\sigma_j^k, \sigma_j^{\tau} \sim \text{Gamma}(1,1).$$

The difference in the regression coefficients between category conditions is modeled directly as $\Delta\beta$:

$$\beta_{bird}^{k} = \Delta\beta^{k} + \beta_{car}^{k},$$

$$\beta_{bird}^{\tau} = \Delta\beta^{\tau} + \beta_{car}^{\tau}.$$

The priors are standard normal distributions (Rouder et al., 2017):

$$\Delta\beta^{k}, \Delta\beta^{\tau}, \beta_{car}^{k}, \beta_{car}^{\tau} \sim \text{Normal}(0,1).$$

The priors on the remaining parameters of the LBA are given by:

$$s_i \sim \text{Normal}(\mu^{s}, \sigma^{s}),$$

$$A_i \sim \text{Normal}(\mu^{A}, \sigma^{A}),$$

$$\mu^{s}, \mu^{A} \sim \text{Normal}(1,1),$$

$$\sigma^{s}, \sigma^{A} \sim \text{Gamma}(1,1).$$

The model was programmed in Stan (Carpenter et al., 2016). We drew 1000 samples from the approximate posterior after convergence of the ADVI procedure (Kucukelbir et al., 2017). Algorithm parameters were the same as those used in the previous fitting procedure.

Modeling Results

We present the model predictions of the EBRW at the group and individual level. It is important to note that the EBRW is more constrained than the LBA. In the LBA, there is a separate accumulator for "old" and "new" responses for each lag for each condition giving it a total of 12 drift rates. By contrast, the EBRW constrains the sum of the drift rates to be one. In addition, these drift rates are not free parameters in the LBA but are constrained by a power law that decays as a function of lag.

Despite these constraints, we show that the EBRW predictions are similar to those of the more general LBA, where the only obvious shortcoming lies in the tails of the RT distributions at lags of 2. This slight cost is at the benefit of the theoretical drift rate decomposition, which further constrains the model and allowed us to investigate changes in underlying cognitive mechanisms that accompany visual expertise. This was our primary focus rather than achieving the best fit possible. To preview, we show that visual expertise is accompanied by increases in memory asymptote, governed by the α parameter, and increases in the distinctiveness of exemplars, governed by the η parameter.

Model Predictions

Panel A of Figure 7 shows the observed and predicted group-level accuracy as function of category and lag. While the model slightly overestimates the overall accuracy, it successfully captures the decrease in hit rates with increases in lag observed in the car condition as well as the similar hit rates across lags in the bird condition. Panel B of Figure 7 shows the predicted individual-level hit rates as a function of the observed hit rates for each lag and category (all $r >$ .90). When compared to the individual-level fit of the baseline LBA in the previous section, the model appears to perform similarly. Panels C and D show that the model is able to accurately capture false alarm rates at both the group and individual level (r = .97 for birds and r = .96 for cars).

Panel A of Figure 8 shows the observed and predicted response time quantiles for targets as a function of category, response type, and lag. The most obvious failure of the model is in the upper tails of the response time distributions for targets at lags of 2. Otherwise, the model captures the overall pattern of the response time distributions for hits. Panel B shows the

predictions for individual-level response time quantiles for misses in the bird condition are noisy due to the low number of observations (r = .59 at lags of 2 and r = .60 at lags of 16), especially at larger quantiles. For the car category, fits to response times for misses are slightly better due to the higher number of observations (r = .78 for lags of 2 and r = .82 for lags of 16). Predictions for response times for hits for both the bird (r = .93 for lags of 2 and r = .92 for lags of 16) and car category (r = .94 for lags of 2 and r = .88 for lags of 16) are much better and the model only slightly overestimates the upper quantiles as was shown in the group-level fits. Panel C shows that the model captures the group-level response time data for foils. Panel D of Figure 8 shows that the model was able to successfully capture individual-level response times for false alarms and correct rejections in both the bird (r = .85 and r = .95, respectively) and car (r = .90 and r = .93, respectively) categories.

Qualitatively, the overall fits for the EBRW model were satisfactory, but not quite as good as those obtained for the more general LBA model in which drift rate was simply a free parameter in every condition; the significant constraints that the EBRW places on the model resulted in a fit that was slightly worse. Quantitatively, correlations between predicted and observed hits and false alarm rates were high, r > .90. We also observed high correlations between predicted and observed response times except for misses, as was the case for the baseline LBA model. Given that the model captures most of the key trends in the data at the group and individual-level, we traded a slightly worse fit for improved theoretical insight into underlying memory processes.

What Causes Drift to Vary with Expertise?

Figure 9 show the predicted parameter estimates as a function of the estimated expertise index. The top row shows the predicted memory asymptote parameter, α, as a function the expertise index, $\Delta d'$. The data provided strong evidence in favor of a category x expertise interaction, $BF = 2.93 \times 10^{-7}$, $\Delta\beta^{\alpha}$ 95% HDI [1.37, 2.50]; the posterior distribution of the regression coefficient on the expertise index for the car category, β^{α}_{car}, had a 95% HDI that fell below zero, [-.31, -.17], where the 95% HDI for the bird category, β^{α}_{bird}, fell well above zero, [1.14, 2.29]. Thus, memory asymptote increased as visual expertise increased for the bird category but not for the car category.

The memory decay parameter, γ, did not show a category x expertise interaction, $BF = 4.59$, $\Delta\beta^{\gamma}$ 95% HDI [-.35, .47]; the 95% HDI of the regression coefficients on the expertise index for the car category, β^{γ}_{car}, and bird category, β^{γ}_{bird}, contained zero, [-.02, .20] and [-.26, .56], respectively. Thus, changes in visual expertise were not accompanied by changes in the rate of decay of exemplar activations.

The similarity between exemplars showed an interaction between category and expertise, $BF = 4.69 \times 10^{-5}$, $\Delta\beta^{\eta}$ 95% HDI [-.37, -.22]; the posterior distribution of the regression weight on the expertise index for the car category, β^{η}_{car}, had a 95% HDI that included zero, [-.07, .004], but the 95% HDI for the bird category, β^{η}_{bird}, fell below zero [-.41, -.23]. Thus, the distinctiveness of exemplars increased as visual expertise increased for the bird category but not for the car category.

There was no category x expertise interaction for response threshold, $BF = 125.83$, $\Delta\beta^{k}$ 95% HDI [-.01, .02]. The posterior distributions of the regression coefficients on the expertise

index for the car category, β^k_{car}, and bird category, β^k_{bird}, had 95% HDI that fell below but close to zero, [-.04, -.01] and [-.04, -.02], respectively. Non-decision time also did not show a category x expertise interaction, $BF = 27.91$, $\Delta\beta^\tau$ 95% HDI [-.002, .03]. The posterior distributions of the regression coefficients on the expertise index for the car category, β^τ_{car}, and bird category, β^τ_{bird}, had 95% HDIs that included zero, [-.01, .03] and [-.02, .003], respectively. Thus, non-decision time did not vary with expertise.

General Discussion

Visual expertise is accompanied by better short-term and long-term recognition memory for images within an expert domain. While these empirical findings replicate past findings of the impact of expertise on short-term (e.g., Curby et al., 2009) and long-term memory (e.g., Herzmann & Curran, 2011), here we examined both within a continuous recognition memory task. We also went beyond past empirical work by comparing cognitive models that instantiated alternative hypotheses about the impact of expertise on memory processes.

We first analyzed continuous recognition memory performance, including both accuracy and response times, using the Linear Ballistic Accumulator model (LBA; Brown & Heathcote, 2008). LBA allowed us to decompose memory performance into three psychological components – perceptual encoding time, response threshold, and drift rate – and measure how these components varied with visual expertise.

The perceptual encoding time parameter is theoretically related to the efficiency of perceptual processing and comparing perceptual representations with memory representations. While there is good reason to think that the development of visual expertise may impact the efficiency of these initial stages of processing during memory tasks (e.g. Curby & Gauthier,

2007; Curby et al., 2009; Gauthier et al., 2000), we did not observe significant variation in perceptual encoding time with expertise. This is certainly not to say that perceptual encoding *mechanisms* do not change with visual expertise. In fact, it is quite likely that they do (e.g., Palmeri & Cottrell, 2009; Palmeri et al., 2004). While the time it takes for perceptual processing may not decrease with expertise (Mack & Palmeri, 2011), at least in the context of recognition memory, the quality of perceptual and memory representations may well increase with expertise, as reflected in the drift rate changes we observed with expertise. The null effect of expertise on non-decision time does not indicate a null effect of expertise on quality of perceptual and memory representations. In fact, our model provides direct evidence for this when we model drift rates directly.

Differences in LBA response threshold reflect differences in response threshold, which affect speed-accuracy tradeoffs. We observed an overall decrease in response threshold with visual expertise, but this was observed for both expert domain and non-expert domain images. Since these differences were not specific to an expert domain, we can only speculate as to what might drive this effect as we did not develop and test a formal model, but individual differences in confidence, motivational factors, or age might be possibilities. For example, expertise might increase confidence in the ability to remember domain-specific information and this confidence might leak over to blocks in the memory experiment containing non-expert domain stimuli, similarly affecting the decreased level of response threshold for non-expert images even at the cost of poorer memory performance. We did find that age was negatively correlated with expertise and was positively correlated with RTs. Therefore, the general threshold increases with expertise may reflect age-related slowing. Prior studies investigating the effects of aging on memory performance have also found increases in threshold with age. Older subjects aged 60-75

have been shown to have more conservative response thresholds compared to college-age subjects for recognition memory (Ratcliff, Thapar, & McKoon, 2004, 2011) and associative memory (Ratcliff et al., 2011).

Drift rate reflects the quality of the evidence upon which a recognition memory decision is based. As previous studies have suggested (Herzmann & Curran, 2011; Lorenc et al., 2014), increases in visual expertise could well increase the quality of memory representations for images within a domain of expertise, which is consistent with our finding greater differences in drift rates for expert-domain versus non-expert-domain images with increases in visual expertise.

To probe further theoretically how visual memory processes might vary with visual expertise, we decomposed drift rate into potential theoretical subcomponents by applying an extension of the EBRW model (Nosofsky, Cox, et al., 2014). This model assumes that when an item is studied, its representation is encoded in long-term memory as an exemplar, and when an item is tested, its representation is matched with stored exemplars. This memory matching results in an activation for each stored exemplar and the summed activation across exemplars provides the drift rate for the LBA. We specifically assumed that memory activation is a joint function of the *strength* with which each exemplar is stored, its *similarity* to the test item, and the amount of *decay* that has occurred since the exemplar was first stored in memory. Thus, the EBRW provides three potential theoretical subcomponents – similarity, memory strength, and decay – that could each vary with visual expertise. We observed significant variation in similarity and memory strength with visual expertise.

Increases in visual expertise were accompanied by decreases in the similarity of the stored exemplars for images in the expert domain. Decreases in similarity with increases in visual expertise could in turn be a result of better memory sensitivity, enhanced exemplar

representations, or more optimal selective attention to relevant dimensions. This increase in distinctiveness may be the result of increases in memory sensitivity, or the quality of the exemplar representations. This is consistent with prior findings for both short- and long-term recognition. In a visual short-term recognition task, Lorenc et al. (2014) found expertise for upright faces increased the quality of short-term memory representations. And in a visual long-term recognition memory task, Herzmann and Curran (2011) found changes in car expertise to be correlated with changes in a parietal event-related potential associated with recollective processes, which depends on robust encoding of distinctive features. In the EBRW and other exemplar-based models, exemplars are represented as vectors of feature values. It may be the case that increases in visual expertise are the result of increases in the probability that features are correctly stored, thereby decreasing the overall amount of memory noise in the system (Nosofsky & Alfonso-Reese, 1999; Shiffrin & Steyvers, 1997), or that these representations become increasingly distant from one another in psychological space as visual expertise increases. Further work is needed to explore the representations in ways we did not do here.

Increases in visual expertise were also accompanied by an increase in memory strength. According to our EBRW-based modeling, memory traces were more strongly encoded for images within the expert domain. When the memory trace is viewed as a vector of feature values, stronger encoding might be realized by increasing the number of dimensions of the memory trace, thereby allowing for the possibility additional features to be stored. For example, birders use beak shape, eye shape, color, size, pattern, etc. to identify the particular species, each of which can be thought of as a dimension in the memory trace vector. As expertise increases, the number of dimensions that a birder can use might also increase. Storing additional features

should, in general, lead to increases in overall memory strength because the summed activation between the cue and the memory trace involves a larger number of summands.

Beyond the specific findings of our modeling work, we demonstrate how the EBRW model can be applied to continuous recognition memory, a task it was not originally designed to do. Nosofsky, Cox, et al. (2014) presented participants with study lists varying in length from 1 to 16 items and showed that both long-term and short-term recognition could be explained via a single exemplar-based model. Extending the Nosofsky, Cox, et al.'s exemplar-based framework as a front-end to the LBA, we found that the model also predicted performance across both long-term and short-term lags, but now in a continuous recognition task. One implication of this successful application is that similar memory processes apply to both standard recognition and continuous recognition. Further research regarding the connection between the two tasks is needed.

More importantly, we demonstrate how the combination of elements from EBRW and LBA can be applied theoretically to online experiments with real-world stimuli testing real-world experts, which often entails fitting models to heterogeneous individual participant data with relatively low numbers of observations per condition. By implementing the models in a Bayesian hierarchical framework, we were able to successfully fit the models to data with far lower numbers observations than is usually available. In hierarchical models, increases in participant sample size not only provide better group-level estimates, but also improved participant-level estimates (Kruschke, 2014). Traditional model fitting often obtains large numbers of observations per condition, either by testing a small number of individual participants for many sessions, or by aggregating over large numbers of participants. Because we neither had large numbers of observations per participant nor aggregated over participants, our model fits are

perhaps quite a bit noisier than what might be expected for cognitive modeling. To be clear, our primary focus was on evaluating theoretical questions concerning how model parameters varied with visual expertise, testing by means of the magnitudes of Bayes Factors, not aiming for observed vs. predicted plots with maximally significant correlations.

Conclusion

Visual expertise has a facilitating effect on visual object recognition memory for expert-domain objects and has been found for both short- and long-term recognition memory performance. Although prior work suggests changes to underlying representations for expert-domain stimuli drives changes in performance, a formal account of this relationship did not exist. In the present work, we used a formal cognitive modeling approach that relates visual expertise and visual recognition memory performance via cognitive processes. Our approach was a two-step process where we first measured how these cognitive processes changed with visual expertise followed by testing why they changed. We recruited participants with varying levels of visual expertise and presented a continuous recognition task designed to measure both short- and long-term recognition memory within a single experiment. We found, for the first time within a cognitive modeling framework, visual expertise to be accompanied by changes in the underlying representations of expert-domain stimuli. In addition, we also demonstrate the capabilities and advantages of the Bayesian hierarchical framework in the context of online experiments with low numbers of trials.

References

Annis, J., Miller, B. J., & Palmeri, T. J. (2017). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods*, *49*, 863–886. https://doi.org/10.3758/s13428-016-0746-9

Annis, J., & Palmeri, T. J. (2018). Bayesian statistical approaches to evaluating cognitive models. *Wiley Interdisciplinary Reviews: Cognitive Science*. https://doi.org/10.1002/wcs.1458

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology*, *44*(2), 310–329. https://doi.org/10.1006/jmps.1998.1249

Baker, J., Côté, J., & Abernehty, B. (2003). Sport-specific practice and the development of expert decision-making in team ball sports. *Journal of Applied Sport Psychology*, *15*, 12–25. https://doi.org/10.1080/10413200390180035

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. *Trends in Cognitive Sciences*, *10*(4), 159–166. https://doi.org/10.1016/j.tics.2006.02.004

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152. https://doi.org/10.1207/s15516709cog0502_2

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *The Quarterly Journal of Experimental Psychology*, *26*(2), 274–284.

Crump, M. J. C., McDonnell, J. V, & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Curby, K. M., & Gauthier, I. (2007). A visual short-term memory advantage for faces. *Psychonomic Bulletin & Review*, *14*(4), 620–628. https://doi.org/10.3758/BF03196811

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 94–107. https://doi.org/10.1097/MPG.0b013e3181a15ae8.Screening

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223.

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151. https://doi.org/10.1016/j.jmp.2010.10.001

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*(6), 1026–1036.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 691–698). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.038

Evans, K. K., Cohen, M. a, Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2011). Does visual expertise improve visual recognition memory? *Attention, Perception, & Psychophysics*, *73*(1), 30–35. https://doi.org/10.3758/s13414-010-0022-5

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience*, *6*(4), 428–432. https://doi.org/10.1038/nn1029

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. https://doi.org/10.1038/72140

Gauthier, I., Tarr, M., & Bub, D. (2010). *Perceptual expertise: Bridging brain and behavior*. New York: Oxford University Press.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93.

Green, D. A., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hagen, S., Vuong, Q. C., Scott, L. S., Curran, T., & Tanaka, J. W. (2014). The role of color in expert object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(9), 1–13. https://doi.org/10.1037/xhp0000139

Herzmann, G., & Curran, T. (2011). Experts' memory: An ERP study of perceptual expertise effects on encoding and recognition. *Memory & Cognition*, *39*(3), 412–432. https://doi.org/10.3758/s13421-010-0036-1

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Burlington: Academic Press.

Kruschke, J. K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. *The Oxford Handbook of Computational and Mathematical Psychology*, 279–299. https://doi.org/10.1093/oxfordhb/9780199957996.013.13

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, *18*, 1–45. https://doi.org/10.3847/0004-637X/819/1/50

La Rouche, G. P. (2006). Birding in the United States: A demographic and economic analysis. In G. C. Boere, C. A. Galbraith, & D. A. Stroud (Eds.), *Waterbirds around the world* (pp. 841–846). Edinburgh, UK: The Stationery Office.

Lee, M. D. (2008a). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*(2), 450–456. https://doi.org/10.3758/BRM.40.2.450

Lee, M. D. (2008b). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15. https://doi.org/10.3758/PBR.15.1.1

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), 376–400. https://doi.org/10.1037/0033-295X.95.4.492

Lorenc, E. S., Pratte, M. S., Angeloni, C. F., & Tong, F. (2014). Expertise for upright faces improves the precision but not the capacity of visual working memory. *Attention, Perception, & Psychophysics*, *76*, 1975–1984. https://doi.org/10.3758/s13414-014-0653-z

Mack, M. L., & Palmeri, T. J. (2011). The timing of visual object categorization. *Frontiers in Psychology*, *2*(JUL), 1–8. https://doi.org/10.3389/fpsyg.2011.00165

McGugin, R. W., & Gauthier, I. (2010). Perceptual expertise with objects predicts another hallmark of face perception. *Journal of Vision*, *10*(4), 1–12. https://doi.org/10.1167/10.4.15

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research*, *69*, 10–22. https://doi.org/10.1016/j.visres.2012.07.014

Nosofsky, R. M., & Alfonso-Reese, L. A. (1999). Effects of similarity and practice on speeded classification response times and accuracies: Further tests of an exemplar-retrieval model. *Memory & Cognition*, *27*(1), 78–93. https://doi.org/10.3758/BF03201215

Nosofsky, R. M., Cao, R., Cox, G. E., & Shiffrin, R. M. (2014). Familiarity and categorization processes in memory search. *Cognitive Psychology*, *75*, 97–129. https://doi.org/10.1016/j.cogpsych.2014.08.003

Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1524–1539. https://doi.org/10.1037/xlm0000015

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

Nosofsky, R. M., & Palmeri, T. J. (2015). An exemplar-based random-walk model of categorization and recognition. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford Handbook of Computational and Mathematical Psychology* (pp. 142–164). New York: Oxford University Press.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2017). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0884-8

Nosofsky, R. M., & Stanton, R. D. (2006). Speeded old-new recognition of multidimensional perceptual stimuli: Modeling performance at the individual-participant and individual-item

levels. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 314–334. https://doi.org/10.1037/0096-1523.32.2.314

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 324–354. https://doi.org/10.1037/0278-7393.23.2.324

Palmeri, T. J., & Cottrell, G. W. (2009). Modeling perceptual expertise. In D. N. Bub, M. J. Tarr, & I. Gauthier (Eds.), *Perceptual expertise: Bridging brain and behavior* (pp. 197–245). New York, NY: Oxford University Press.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 309–328. https://doi.org/http://dx.doi.org/10.1037/0278-7393.19.2.309

Palmeri, T. J., Wong, A. C. N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences*, *8*(8), 378–386. https://doi.org/10.1016/j.tics.2004.06.001

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice-Decisions.

Psychological Science, *9*(5), 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1016/j.pestbp.2011.02.012.Investigations

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424. https://doi.org/10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464.

Reimers, S., & Maylor, E. A. (2005). Task switching across the life span: Effects of age on general and specific switch costs. *Developmental Psychology*, *41*(4), 661.

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, *39*(3), 365–370.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304.

Shen, J., Mack, M. L., & Palmeri, T. J. (2014). Studying real-world perceptual expertise. *Frontiers in Psychology*, *5*, 857. https://doi.org/10.3389/fpsyg.2014.00857

Shepard, R. N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology*, *62*(3), 302.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review*.

https://doi.org/10.3758/BF03209391

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*(2), 283–317. https://doi.org/10.1037/a0015156

Tanaka, J. W., & Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science*, *12*(1), 43–47.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic-level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482. https://doi.org/10.1016/0010-0285(91)90016-H

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384. https://doi.org/10.1037/a0032222

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the SavageDickey density ratio. *Computational Statistics and Data Analysis*, *54*(9), 2094–2102. https://doi.org/10.1016/j.csda.2010.03.016

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, *13*(3), 33. https://doi.org/10.1167/13.3.33

Wong, Y. K., & Gauthier, I. (2010). Holistic processing of musical notation: Dissociating failures of selective attention in experts and novices. *Cognitive and Affective Behavioral Neuroscience*, *10*(4), 541–551. https://doi.org/10.3758/CABN.10.4.541.Holistic

Notes

[1] In this article, we have chosen to focus our discussion on interpreting differences in the non-decision time parameter with expertise as due to differences in perceptual processing time (perceptual efficiency); even though we focus on "perceptual" differences, we acknowledge an important caveat that perceptual and motor times are simply additive and differences in the parameter could reflect differences in motor execution time.

[2] Similarity is often derived from a multidimensional scaling solution based on similarity ratings between all stimulus pairs. Because of the large number of stimuli required for the continuous recognition memory task, obtaining pairwise similarity ratings was not feasible. We note there has been some advancements towards this end recently (e.g., Nosofsky, Sanders, Meagher, & Douglas, 2017), but these techniques are beyond the scope of the present work.

Acknowledgements

This work was supported by NSF SBE-1257098 and SMA-1640681, the Temporal Dynamics of Learning Center (NSF SMA-1041755), the Vanderbilt Vision Research Center (NEI P30-EY008126), a Discovery Grant from Vanderbilt University, and a training grant from the NIH (NEI T32-EY007135).

Appendix

Panel A of Figure A1 shows hit rates plotted as a function of lag for each category. A two-way repeated measures ANCOVA with expertise and age as covariates revealed hit rates were significantly greater in the bird condition (M = .89, SD = .14) than in the car category (M = .79, SD = .12), $F(1,51) = 40.95$, $p < .0001$, and were significantly greater for lags of 2 (M = .86, SD = .09) than for lags of 16 (M = .82, SD = .11), $F(1,51) = 18.65$, $p < .0001$. The main effects are qualified by a significant category x lag interaction, $F(1,51) = 8.63$, $p < .01$. A simple effects analysis revealed hit rates for the bird category did not significantly differ between lags of 2 (M = .90, SD = .10) and lags of 16 (M = .88, SD = .13), $t(53) = 1.64$, $p = .108$, but for the car category, hit rates were significantly greater for lags of 2 (M = .82, SD = .11) than for lags of 16 (M = .75, SD = .14), $t(53) = 4.59$, $p < .001$. Panel B of Figure A1 shows the hit rates plotted as a function of expertise and lag for each category. The main effect of expertise on hit rates was not significant, $F(1,51) = .04$, $p = .84$. There was a significant expertise x category interaction, $F(1,51) = 7.44$, $p < .01$. However, expertise did not predict hit rates for birds ($\beta = .03$, $p = .09$, adjusted $R^2 = .04$) or cars ($\beta = -.03$, $p = .054$, adjusted $R^2 = .05$). Therefore, the influence of visual expertise on hit rates varied between the bird and car categories, but expertise was not predictive of hit rates in either category.

Panel C of Figure A1 show the false alarm rates for each category. A one-way (category: birds vs. cars; covariates: expertise, age) repeated measures ANCOVA revealed false alarm rates for the bird category (M = .20, SD = .15) were significantly lower than for car images (M = .32, SD = .12), $F(1,51) = 51.79$, $p < .0001$. Panel D of Figure A1 shows there was a significant main effect of expertise on false alarms, $F(1,52) = 4.45$, $p < .05$, such that false alarms decreased with increases in expertise, ($\beta = -.04$, $p < .05$, adjusted $R^2 = .07$). The expertise x category interaction

for false alarm rates was not significant, $F(1,52) = .07$, $p = .792$. Thus, expertise was accompanied by a decrease in false alarms, but this decrease did not significantly differ between the bird and car categories.

Response time (RT). A two-way ANCOVA with expertise and age as covariates was conducted on median RTs for hits. Panel A of Figure A2 shows median RTs for hits were significantly faster for lags of 2 ($M = 864.92$, $SD = 125.96$) than for lags of 16 ($M = 965.93$, $SD = 134.66$), $F(1,51) = 84.57$, $p < .0001$, but did not significantly differ between category, $F(1,52) = .24$, $p = .628$. The category x lag interaction was not significant, $F(1,52) = 1.20$, $p = .28$. Panel B of Figure A2 shows there was no significant main effect of expertise on median RTs for hits, $F(1,52) = 1.81$, $p = .04$ and no expertise x category interaction, $F(1,52) = .00$, $p = .963$, or lag, $F(1,52) = .19$, $p = .662$. Thus, expertise was not accompanied by changes in response time for hits.

A two-way ANCOVA with expertise and age as covariates was conducted on median RTs for misses with six participants removed from the analysis due to not having missed any targets in at least one condition. Panel A of Figure A2 shows there was no significant main effect of category, $F(1,45) = .05$, $p = .824$, or lag on median RTs for misses, $F(1,45) = 1.39$, $p = .244$. The category x lag interaction was not significant, $F(1,45) = 1.07$, $p = .306$. Panel B of Figure A2 shows there was no significant main effect of expertise, $F(1,45) = .21$, $p = .650$, no significant expertise x category interaction, $F(1,45) = 1.83$, $p = .2183$, and no significant expertise x lag interaction, $F(1,45) = .54$, $p = .467$. Thus, changes in visual expertise were not accompanied by changes in response times for misses.

A one-way (category: bird vs. car) ANCOVA with expertise and age as covariates was conducted on median RTs for false alarms. Panel C of Figure A2 shows median RTs for false

alarms were significantly slower in the bird condition (M = 1081.62, SD = 191.77) than in the car condition (M = 1049.08, SD = 223.23), $F(1,51) = 14.53$, $p < .001$. Panel D of Figure A2 shows there was no significant main effect of expertise, $F(1,52) = 1.23$, $p = .273$, and no significant expertise x category interaction, $F(1,52) = .07$, $p = .800$. Thus, there were group-level effects in which RTs for false alarms were slower in the bird condition, but at the individual-level, expertise was not associated with changes in RTs for false alarms.

Panel C of Figure A2 shows median RTs for correct rejections did not significantly differ between conditions, $F(1,51) = 2.09$, $p = .154$. Panel D shows there was a main effect of expertise, $F(1,51) = 5.83$, $p < .05$, but no interaction, $F(1,51) = .41$, $p = .52$. A multiple regression analysis with age and expertise as predictors revealed that median correct rejection RTs decreased with increased expertise, ($\beta = -64.23$, $p < .05$) and RTs increased with increased age ($\beta = 3.68$, $p < .05$). Because visual expertise was negatively correlated with age, $r(51) = -.30$, $p < .05$, decreases in correct rejection RTs may be due to age-related slowing.

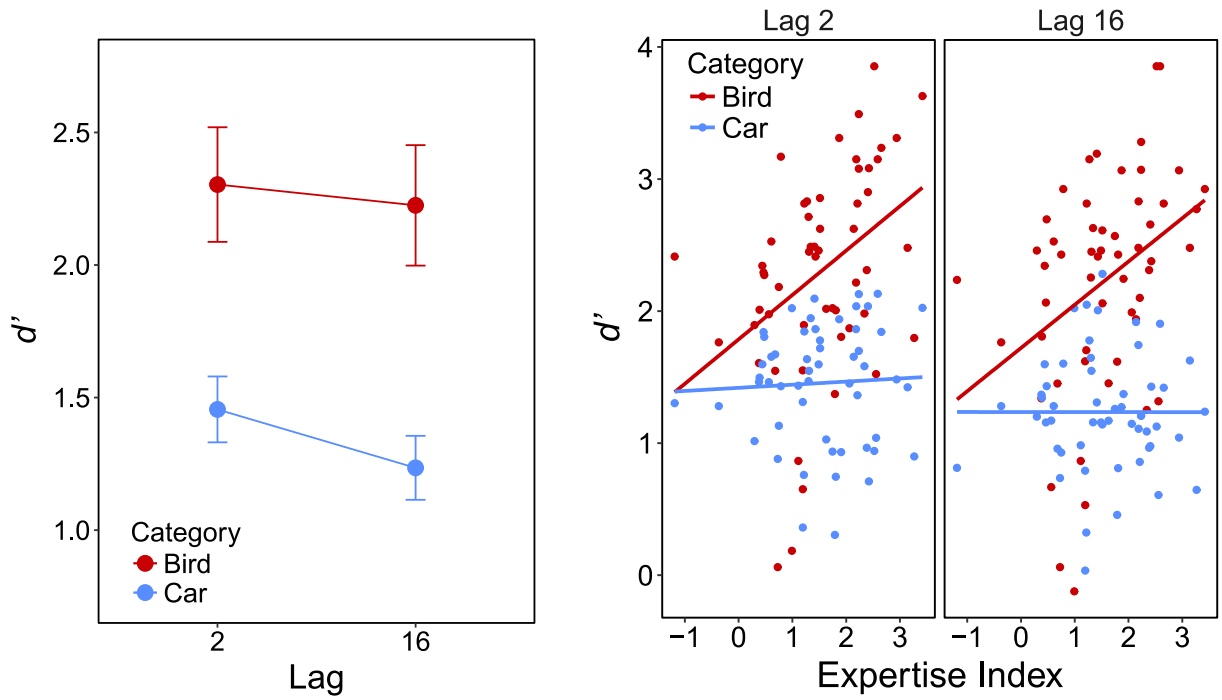

Figure 1. The left panel plots *d*' as a function of lag and category. The right panel shows *d'* for each participant as a function of expertise index and lag for each category.

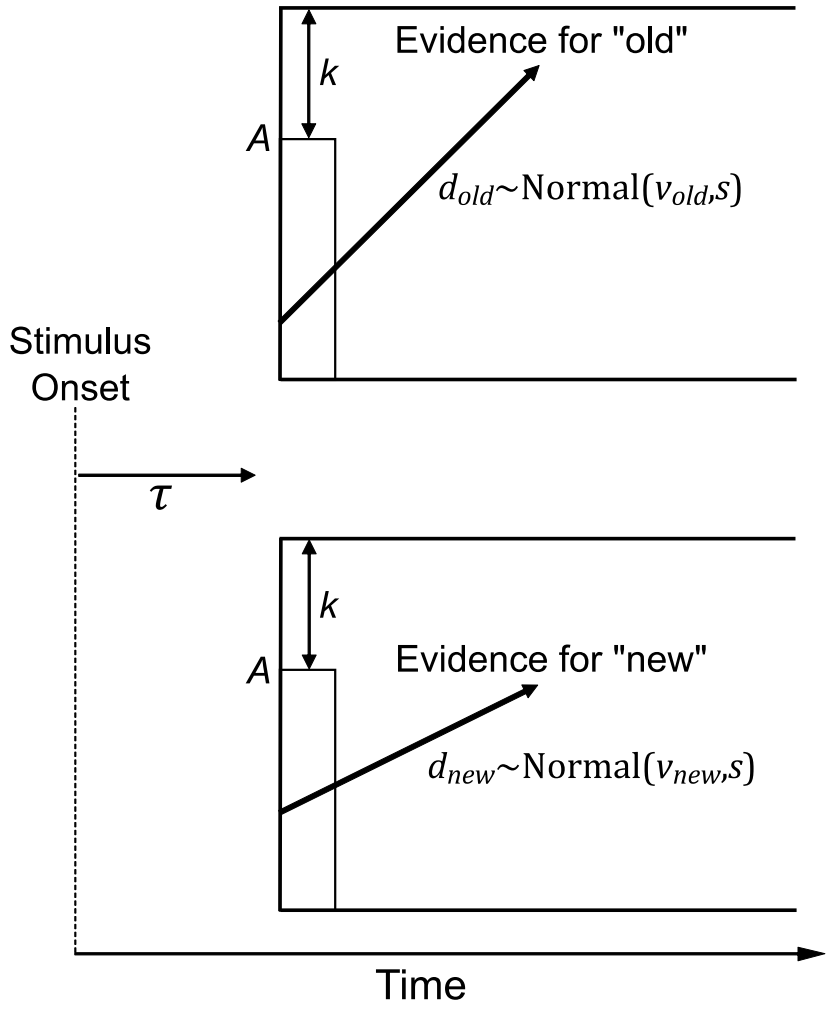

Figure 2. The Linear Ballistic Accumulator model. After stimulus presentation, the stimulus is perceptually encoded with time τ, and evidence begins to accumulate towards either the *old* or *new* response. The rate at which evidence accumulates for each response type is given by the drift rates, d_{old} and d_{new}. The drift rates are assumed to be drawn from corresponding normal distribution with mean v_{old} or v_{new} and standard deviation *s*. The LBA assumes that the starting point of each accumulator varies from trial to trial, where the starting point is drawn from a uniform distribution from 0 to A. The response threshold is given by $A + k$, where *k* is referred to as the relative threshold. When an accumulation process reaches its threshold, the corresponding old or new response is made.

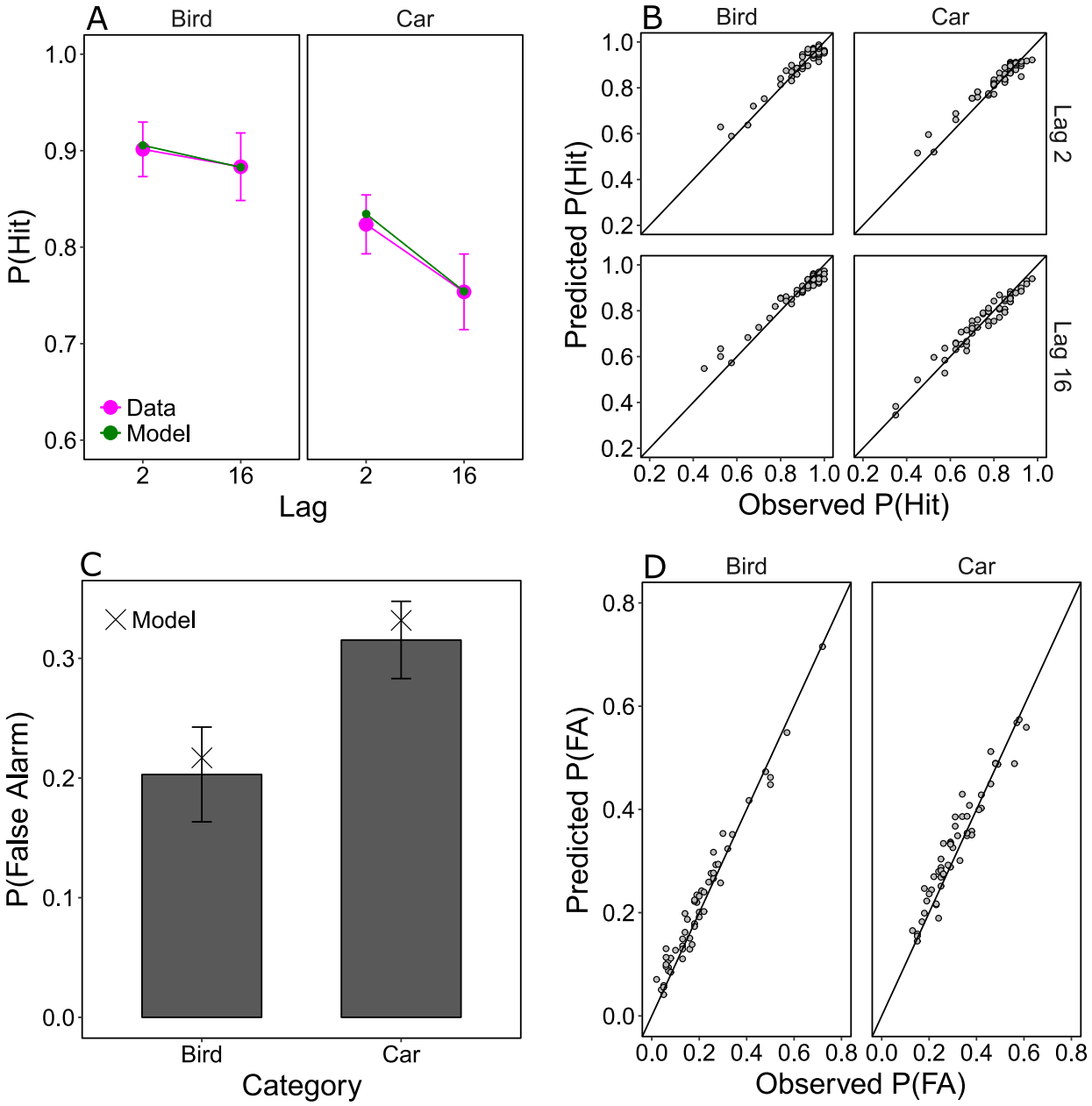

Figure 3. Panel A shows group-level fits of the LBA to the hit rates as a function of lag. Panel B shows predicted hit rates plotted as a function observed hit rates for each participant. Panel C shows fits of the model to false alarm rates for each category. Panel D shows the predicted false alarm rates as a function of the observed false alarm rates for each participant.

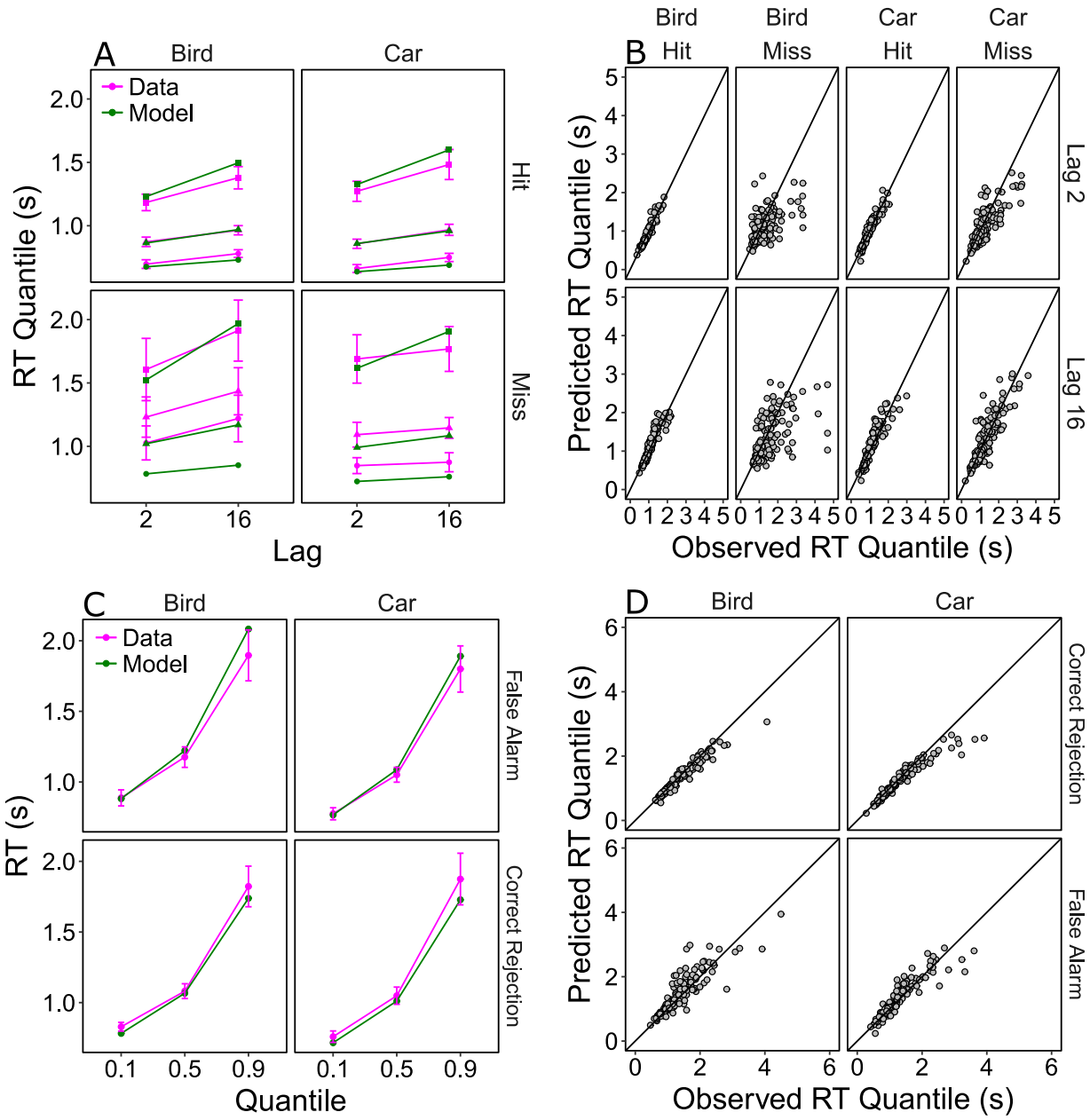

Figure 4. Panel A shows group-level fits of the LBA to the hit rates as a function of lag. Panel B shows predicted hit rates plotted as a function observed hit rates for each participant. Panel C shows fits of the model to false alarm rates for each category. Panel D shows the predicted false alarm rates as a function of the observed false alarm rates for each participant.

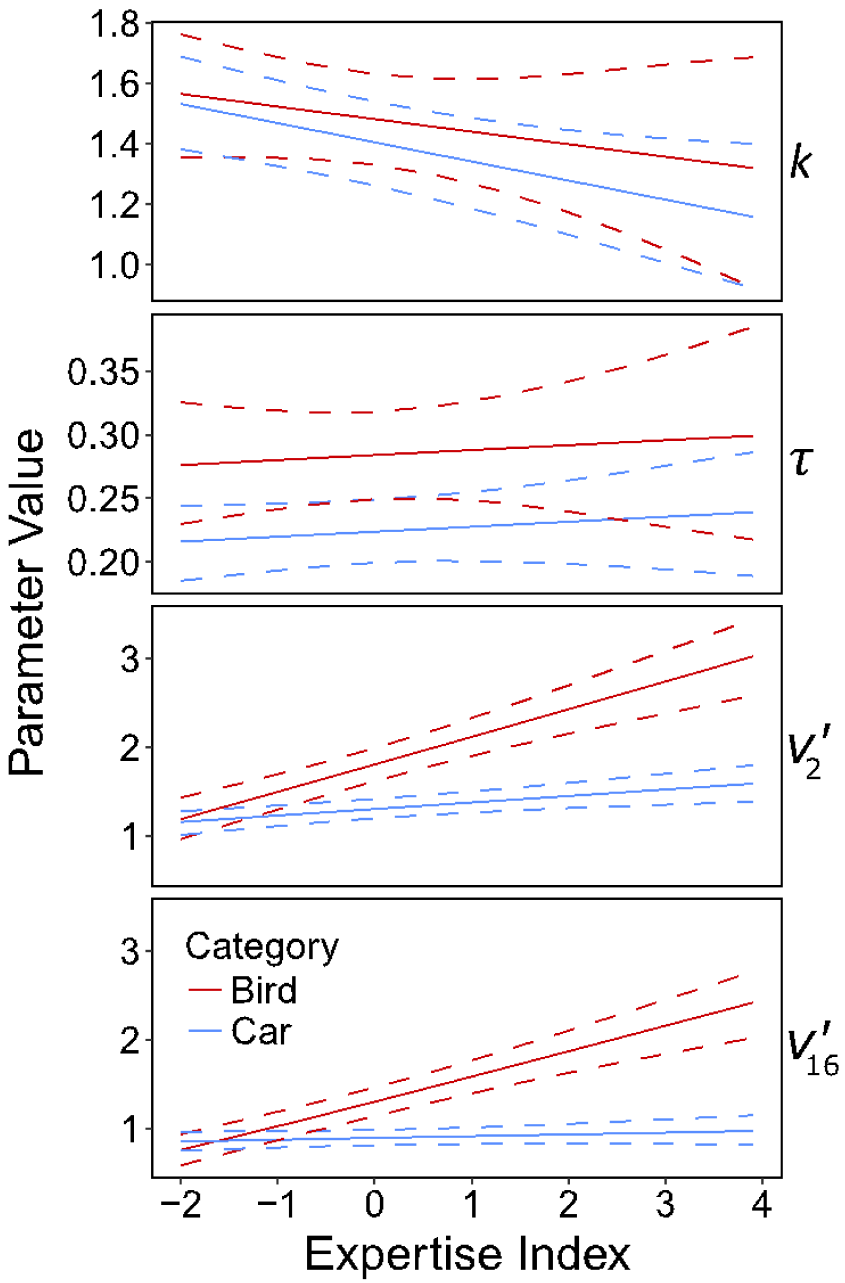

Figure 5. Solid lines show the mean posterior predicted value for each key parameter of the LBA as a function of the expertise index for each category. Dotted lines show the 95% highest density interval. Parameter meanings: k: response threshold; τ: non-decision time; v_2': difference in drift rate between new items and old items at lags of 2; v_{16}': difference in drift rate between new items and old items at lags of 16.

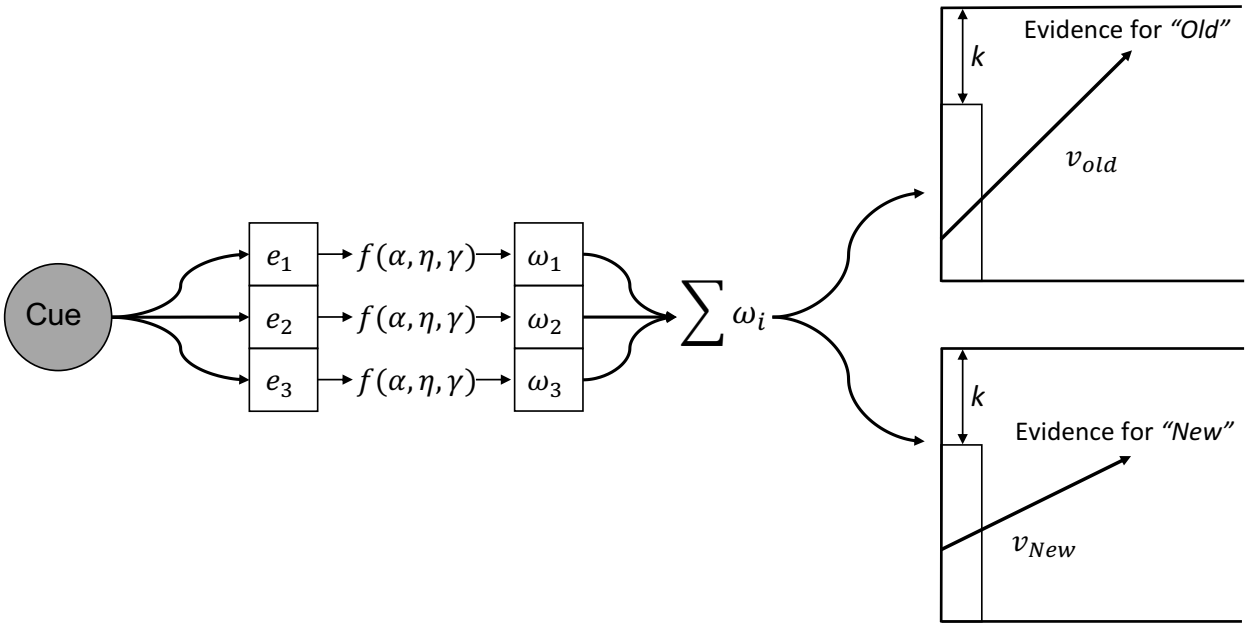

Figure 6. A graphical representation of the EBRW model for continuous recognition. The cue is matched in parallel to the activated set of stored exemplars, shown here as $e_1, e_2,$ and e_3. The matching process results in an activation value for each exemplar, shown here as ω_1, ω_2 and ω_3. These activations are a joint function of the overall memory strength of the stored exemplar, α, the similarity between the cue and exemplar, *s*, and the amount of memory decay that has occurred since storage, γ. These activations are then summed and normalized, which is then used as the mean drift rate in the accumulator for the "old" response. The mean of the new drifts rate is computed by subtracting the sum of the activations from 1.

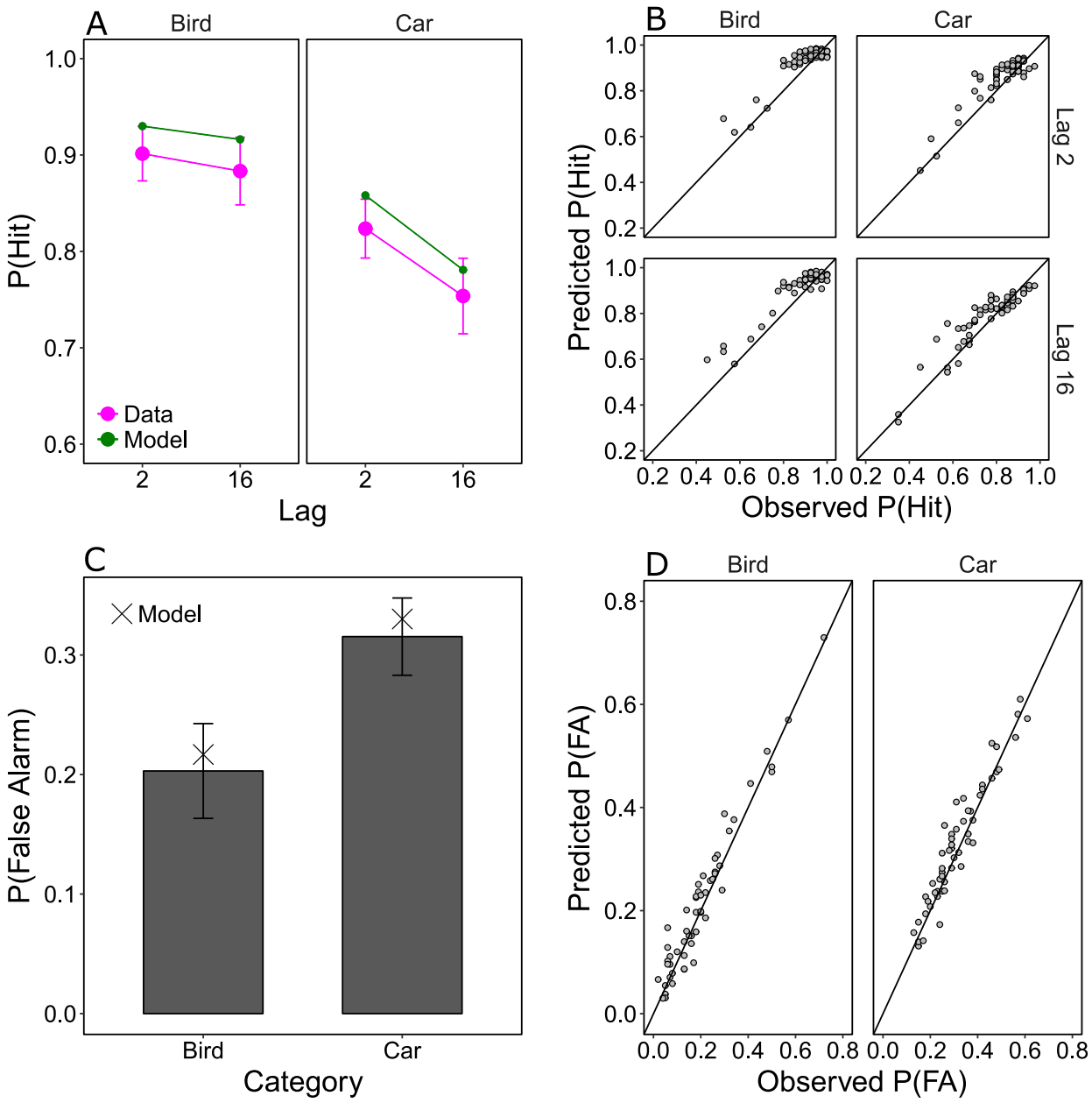

Figure 7. Panel A shows group-level fits of the EBRW to the hit rates as a function of lag. Panel B shows predicted hit rates plotted as a function observed hit rates for each participant. Panel C shows fits of the model to false alarm rates for each category. Panel D shows the predicted false alarm rates as a function of the observed false alarm rates for each participant.

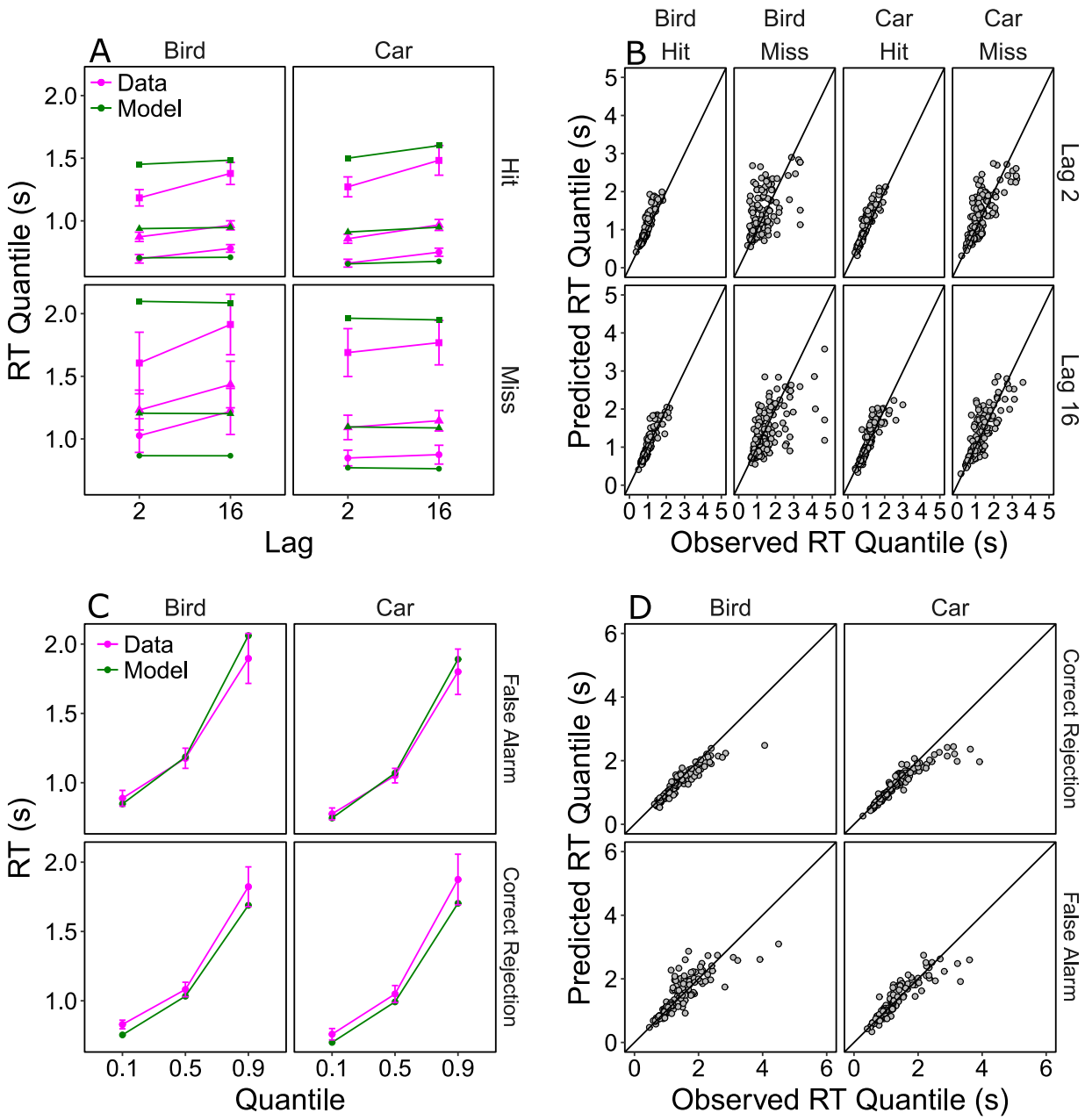

Figure 8. Panel A shows group-level fits of the EBRW to hit and miss RT quantiles (10%, 50%, 90%) as a function of lag. Quantiles Panel B shows predicted hit and miss RTs plotted as a function observed hit rates for each participant. Panel C shows fits of the model to false alarm and correct rejection RT quantiles for each category. Panel D shows the predicted individual-level RT quantiles as a function of the observed quantile for each participant.

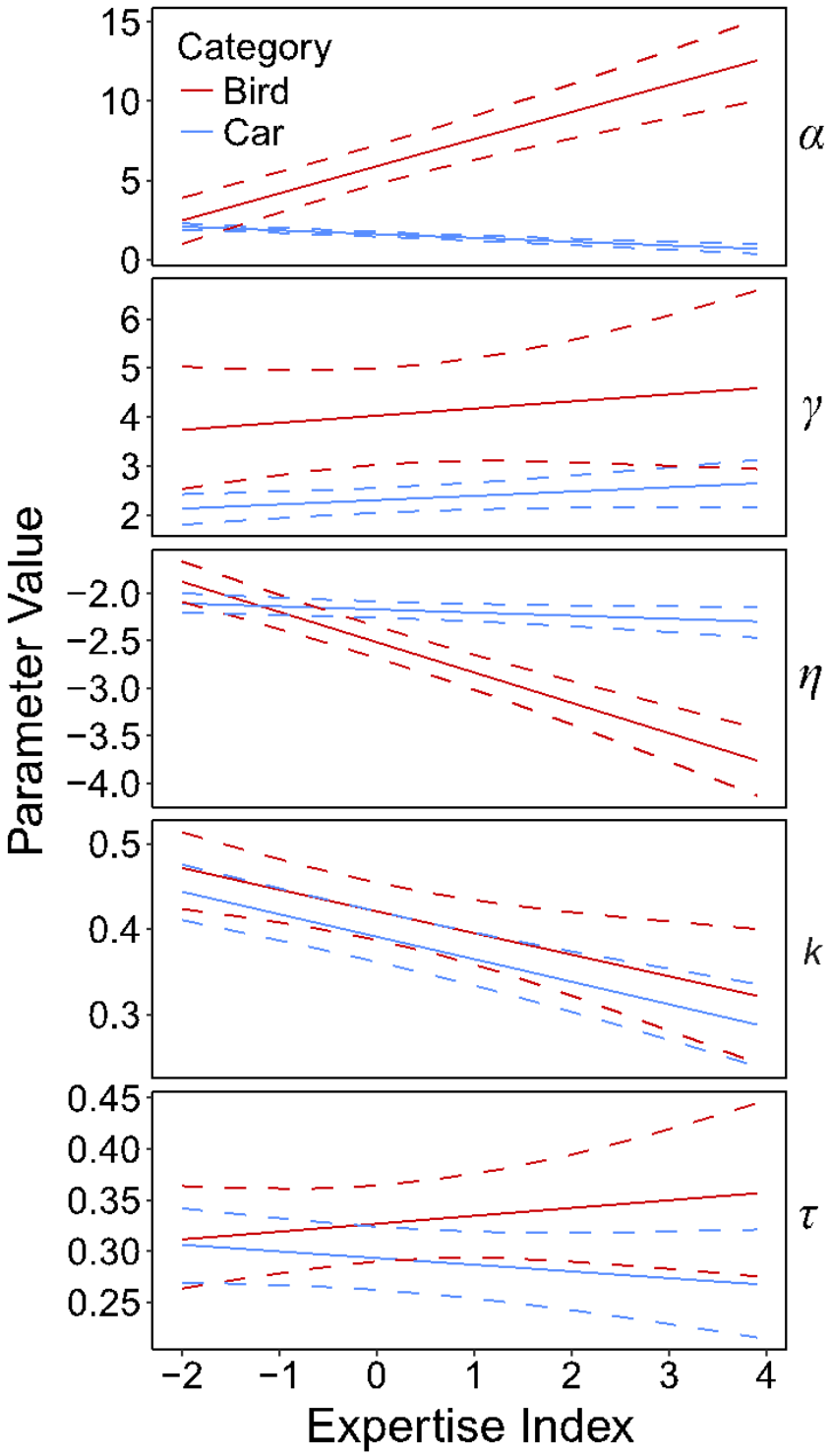

Figure 9. Solid lines show the mean posterior predicted value for each key parameter of the EBRW as a function of the expertise index for each category. Dotted lines show the 95% highest density interval. Parameter meanings: α: overall memory strength; γ: memory decay; η: similarity, k: response threshold; τ: non-decision time.

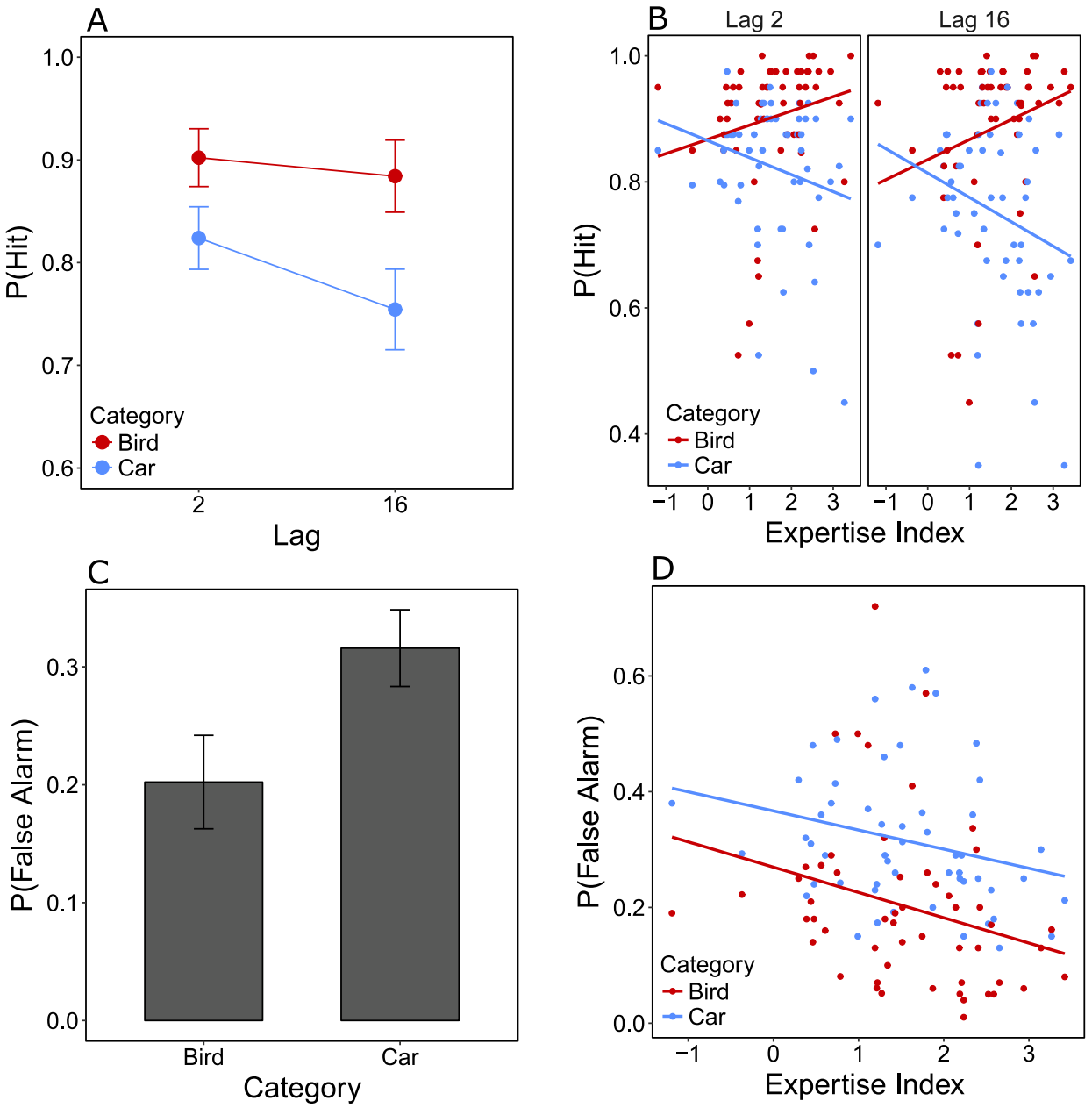

Figure A1. Panel A shows the hit rate as a function of lag and category. Panel B shows the individual-level hit rates as a function of expertise index and lag for each category with simple linear regression lines. Panel C shows the false alarm rates as a function category and Panel D shows the false alarm rates as a function of expertise index and category with regression lines.

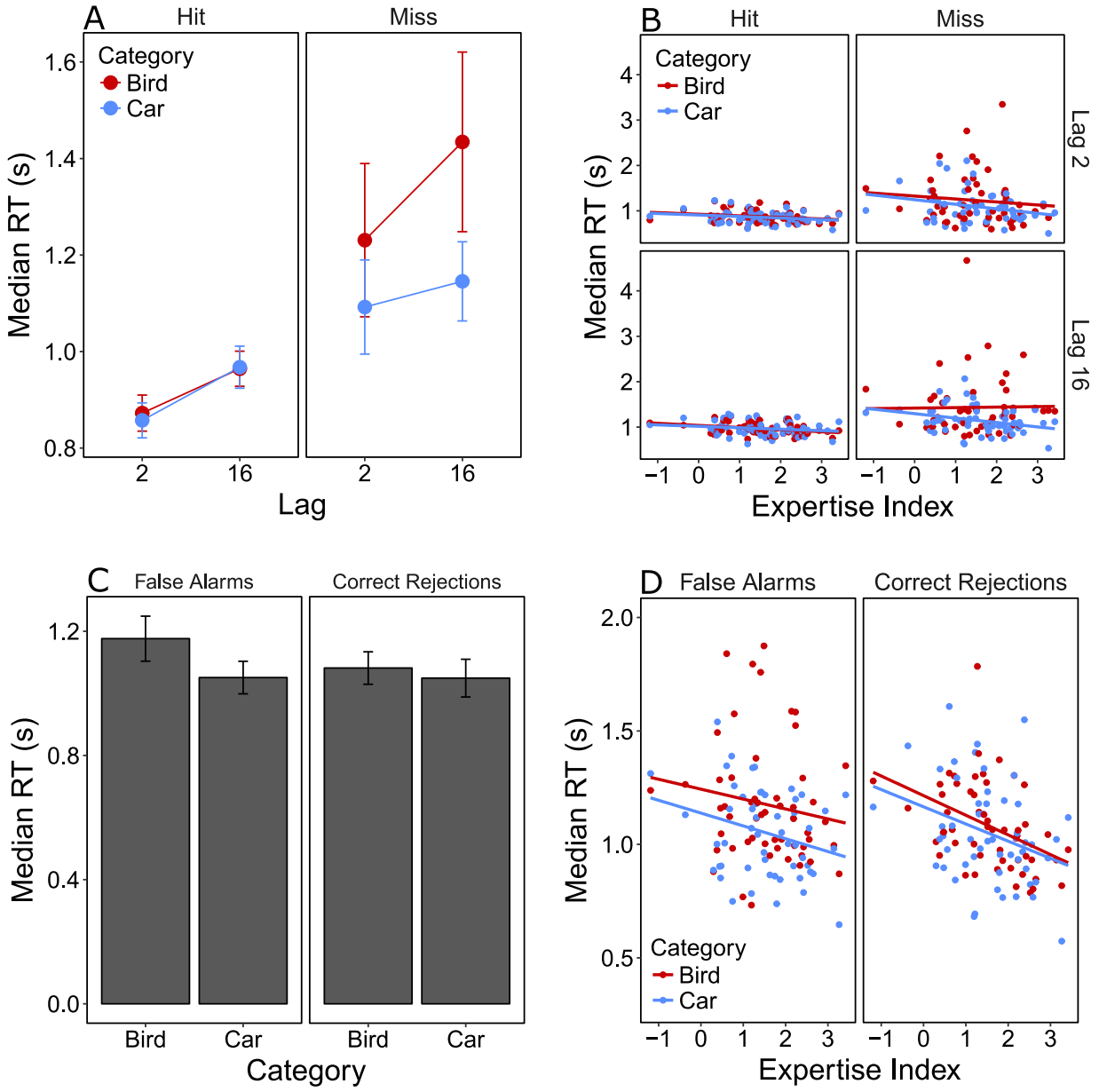

Figure A2. Panel A shows the median hit and miss RTs as a function lag and category. Panel B shows the participant-level median hit and miss RTs as function of their expertise index for each category and lag. Lines represent simple linear regression lines. Panel C shows the median false alarm and correct rejection RTs as a function of category. Panel D shows the participant-level false alarm and correct rejections as a function of their expertise index for each category.