How to Make a Proceedings Paper Submission

Anonymous CogSci submission

Abstract

Include no author information in the initial submission, to facilitate blind review. The abstract should be one paragraph, indented 1/8 inch on both sides, in 9~point font with single spacing. The heading 'Abstract' should be 10~point, bold, centered, with one line of space below it. This one-paragraph abstract section is required only for standard six page proceedings papers. Following the abstract should be a blank line, followed by the header 'Keywords' and a list of descriptive keywords separated by semicolons, all in 9~point font, as shown below.

Keywords: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

Introduction

Whether to keep looking at a current target of attention is one of the most fundamental decisions we make, whether we are trying to find our way in a busy street or swiping through TikTok. Even young infants constantly decide what to look at and for how long. In fact, infant research has capitalized on infants' ability to endogenously control their attention. Through the use of looking time paradigms, researchers have been able to make inferences about infants' learning and mental representations from the changes in their looking duration(Aslin, 2007; Sim & Xu, 2019). In a typical experiment, infants decrease their looking duration upon seeing repeated stimulus (i.e. habituation). After being habituated, infants' interest will often recover when seeing a novel stimulus, relative to the habituation stimulus (i.e. dishabituation). While these phenomena are well-documented, the mechanisms underlying them remain poorly understood. A better understanding of what shapes habituation and dishabituation is of both methodological and theoretical significance. Methodologically, it is central to relating behavior to infants' internal representations. Theoretically, it would shed light on infants' active role in shaping their own learning [Smith, Jayaraman, Clerkin, & Yu (2018); raz2020learning] and reveal principles that guide human information-seeking behavior in general.

Classical theory of infant looking behavior suggests three factors are crucial to habituation and dishabituation: complexity, familiarization time, and infants' age (Hunter & Ames, 1988): 1) Infants will take longer to habituation to complex stimuli, 2) longer familiarization time to one stimulus would make infants more likely to dishabituate to another stimulus and 3) The older the infants, the faster their information processing, and the faster they will habituate. Together, these three factors determine how infants' looking time changes during an experiment. Although this theory is influential, few empirical work has examined these three factors systematically (for exception, see Hunter, Ames, & Koopman, 1983) and the lack of quantitative details in the theory has made it impossible to offer precise predictions.

In contrast to the classical verbal theory, recent work has attempted to describe infants' looking behaviors through computational modeling. In pioneering work, Kidd, Piantadosi, & Aslin (2012) developed a paradigm in which infants are shown sequences of events. Infants' look-away probabilities toward the stimuli are compared with surprisal, a measure of information content, derived from a rational learner model that keeps track of the probabilities of each event. The study shows that infants' pay most attention to event sequences that are neither too high nor too low in surprisal, resulting in a 'Golidlocks' effect of attention. A recent study by Poli, Serino, Mars, & Hunnius (2020) offered an alternative linking hypothesis between the model and behavior: the study used a similar paradigm and model to show that infants' looking time can be predicted by the Kullback-Leibler (KL) divergence. KL measures the statistical distance between two probabilities distribution, which intuitively tracks 'learning progress.' These attempts on connecting information theoretic measurements to infants' looking time resonate with the emerging literature on curiosity in developmental robotics and reinforcement learning (Haber, Mrowca, Fei-Fei, & Yamins, 2018; Oudeyer, Kaplan, & Hafner, 2007). Curiosity-driven artificial agents' exploratory behaviors are guided by optimizing expected information gain (EIG), a measurement that has been shown to be related to curiosity-driven learning iin human children and adults (Liquin, Callaway, & Lombrozo, 2021).

However, there are several limitations to the existing models. First, current models do not capture the noisy nature of perceptual learning (Callaway, Rangel, & Griffiths, 2021; Kersten, Mamassian, & Yuille, 2004). That is, the models were assumed to acquire perfect representation of each event in the sequence. This assumption leads to the second limitation: While surprisal and KL-divergence have been shown to correlate with infants' looking behaviors, current models do not provide an account of how these measurements might be linked mechanistically to the infants' behavior. That is, while these models show that infants might be sensitive to the information-theoretic variability in their learning environment, they do not provide an account of how they are related to infants' real-time sampling behavior. Finally, the event sequence paradigm used to evaluate these models are not representative of classical infant looking time paradigms. As the key phenomena described in the Hunter & Ames (1988) theory were not captured, the extent to which we can extrapolate current behavior-model fits to behavior in a typical looking time experiment remains limited.

Here we present steps to overcome these limitations. Our goal is to provide a unifying quantitative account of looking behaviors as arising from optimal decision-making over noisy perceptual representations (Bitzer, Park, Blankenburg, & Kiebel, 2014; Callaway et al., 2021). We begin by instantiating a version of prior learning models in an independent-trial format (where individual stimuli are learned, not sequences of events). We then develop a second model that addresses weakness in previous work by presenting a model that a) accumulates noisy samples from the stimulus, and b) directly chooses what to look at using different information-theoretic linking hypotheses (surprisal, KL-divergence, and EIG). Finally, we evaluate our model with adult looking time data collected from a paradigm that captures habituation, dishabituation, and complexity effects.

Models

We reasoned that, at its simplest, habituation occurs when each repetition of a stimulus refines the representation of a concept until repetitions become ineffectual. Dishabituation then occurs when a stimulus deviates from the concept learned during habituation. We therefore formalized the learning problem that participants face in a simple habituation experiment as a form of Bayesian concept learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Tenenbaum, 1999). The goal is to learn a concept θ, which is a set of probabilities for independent binary features $\theta_{1,2,..,n}$, where n is the number of features.

Model 1: Discrete Time Model

In the first model, we assume that the learner receives information about θ by observing exemplars y: instantiations of $\bar{\theta}$, where each feature $y_{1,2,..,n}$ is either on or off. Each feature θ_i and its corresponding exemplar y_i form a Beta-Bernoulli process:

$$p(\theta_i) \sim Beta(\alpha_i, \beta_i) \quad (1)$$

$$p(y_i|\theta_i) \sim Bernoulli(\theta_i) \quad (2)$$

Since the features are independent, this relationship holds for the entire concept θ. In previous work, two information-theoretic quantities, surprisal and Kullback-Leibler (KL) divergence, resulting from the stimulus were shown to be linked to looking behavior (Kidd et al., 2012; Poli et al., 2020). Surprisal, calculated as $-log(p(y|\theta))$, intuitively refers to how surprising a stimulus y is given the model's beliefs about θ - the intuition that surprising events should result in longer looking times has served as a foundational assumption in developmental psychology. KL-divergence measures how much a model needs to change to accommodate a new stimulus y, and describes a distance between the model before and after an observation, in is defined in our case as $\sum_{x \in X} p(\theta = x|y) \frac{p(\theta=x|y)}{p(\theta=x)}$. If an observation causes a large change, we speculated that a proportionally long looking time is necesssary to integrate the new information.

Model 2: Continuous Time Model

A limitation of Model 1 is that it assumes that stimuli are encoded perfectly and instantaneously. However, to model the precise time course of attention, we instead suggest that participants gather repeated noisy samples $\bar{z}$ from the exemplars, instead of directly observing them. For any sample z from an exemplar y there is a small probability ε to misperceive the feature as off when it was actually on, and vice versa.

Therefore, by making noisy observations $\bar{z}$, the learner obtains information about the true identity of the exemplar y, and by extension, about the concept $th\bar{e}ta$. By Bayes' rule:

$$P(\theta|\bar{z}) \quad = p(\bar{z}|y)p(y|\theta)p(\theta)/p(\bar{z}) \quad (3)$$

where $p(\bar{z}|y_i)$ is fully described by ε, and $p(y|\theta)$ by Bernoulli processes as in Eq. 2.

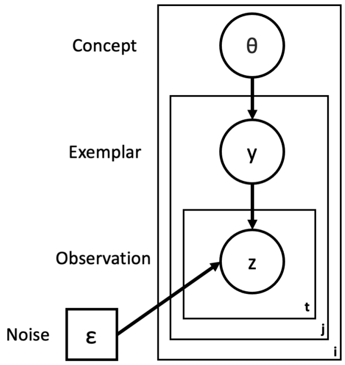

Figure 1: Graphical representation of our model. Circles indicate random variables. The squares indicate fixed model parameters.

Sampling The formulation of the model in continuous time allows us to do two things: First, we can explicitly model the learner's decision on when to stop sampling by asking the model to decide, after every sample z, whether it wants to continue sampling from the same stimulus or not. This is in contrast to the discrete time models presented here and in previous work (Kidd et al., 2012; Poli et al., 2020), where we can only link information-theoretic measures to looking data, but not provide a mechanism for how these measures could control moment-to-moment sampling decisions. Second, a consequence of making a decision at every time step is that we can study the behavior of another information-theoretic measure: the expected information gain (EIG). EIG is commonly used in rational analyses of information-seeking behavior - that is to assess whether information-seeking is optimal with respect to the learning task (Markant & Gureckis, 2012; Oaksford & Chater, 1994). Importantly, EIG is a forward-looking measure that considers the potential for learning from the next sample. Since discrete time models operate on the level of a whole stimulus, rather than individual samples, EIG would look forward to the next stimulus in these models, rather than the next sample, and therefore not be able to capture the decision of whether to keep looking. EIG to describe looking time is therefore only possible in the continuous time models.

We compute EIG by weighing the information gain from each possible next observation by the probability of that observation. We defined information gain as the KL-divergence between the hypothetical posterior after observing a future sample z_{t+1} and the current posterior:

$$EIG(z_{t+1}) = \sum_{z_{t+1}\in[0,1]} p(z_{t+1}|\theta_t) * D_{KL}(\theta_{t+1}||p(\theta_t)) \tag{4}$$

Finally, to get actual sampling behavior from the model, it has to convert EIG into a binary decision about whether continue looking at the current sample, or to advance to the next trial. The model does so using a luce choice between the EIG from the next sample and a constant EIG from looking away.

$$p(look) = \frac{EIG(z_{t+1})}{EIG(z_{t+1}) + EIG(world)} \tag{5}$$

We also studied the behavior of the model when replacing EIG with continuous time versions of the linking hypotheses described earlier, surprisal and KL-divergence between the posterior $p(\theta_t)$ and the prior $p(\theta_{t-1})$.

Experiment

To evaluate how well these models can explain looking time changes, we developed a stimuli set and an experimental paradigm to reproduce the key looking time patterns in adult participants. There are two advantages for evaluating models with adult looking time data: 1) it is relatively easy to acquire adult looking time data to reach sufficient power; 2) adult looking time data can speak to the developmental continuities of the principles guiding looking time behaviors.

Methods

Stimuli We created the animated creatures using Spore (a game developed by Maxis in 2008). There were forty creatures in total, half of which have low perceptual complexity (e.g. the creatures do not have limbs, additional body parts, facial features, or textured skin), and half of which have high perceptual complexity (i.e. they do have the aforementioned features; see Fig.2 for examples). We used the "animated avatar" function in Spore to capture the creatures in motion.

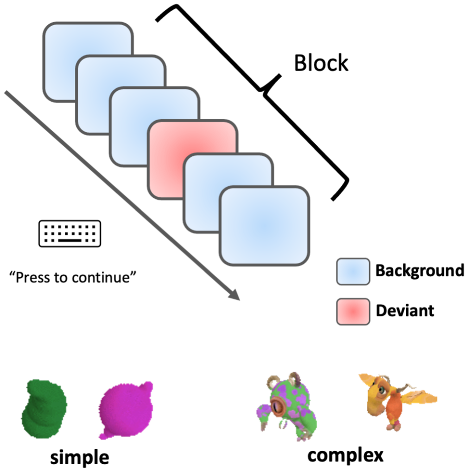

Figure 2: Experimental design and examples of simple and complex stimuli. In each block, a deviant could appear on the second, fourth (as depicted here) or sixth trial or not at all. Stimuli within a block were either all simple or all complex.

Procedure The experiment was a web-based, self-paced visual presentation task. Participants were instructed to look at a sequence of animated creatures at their own pace and answer some questions throughout. On each trial, an animated creature showed up on the screen. participants can press the down arrow to go to the next trial whenever they want after a minimum viewing time of 500 ms.

Each block consisted of six trials. A trial can be either a background trial (B) or a deviant trial (D). A background trial presented a creature repeatedly, and the deviant trial presented a different creature from the background trial in the block. Two creatures in the blocks were matched for visual complexity. There were four sequences of background trials and deviant trials. Each sequence appeared twice, once with high complexity stimuli and once with low complexity stimuli. The deviant trial can appear at either the second (BDBBBB), the fourth (BBBDBB), or the sixth trial (BBBBBD) in the block. Two blocks do not have deviant trials (BBBBBB). The creatures presented in the deviant trials and background trials were matched for complexity. Each participant saw eight blocks in total, half of which used creatures with high perceptual complexity, and half of which used creatures with low perceptual complexity.

To test whether behavior was related to task demands, participants were randomly assigned to one of the three attention check conditions, differing in the type of questions asked following each block: Curiosity, Memory, and Math. In the Curiosity condition, participants were asked to rate "How curious are you about the creature?" on a 5-point Likert scale. In the Memory condition, a forced-choice recognition question followed each block ("Have you seen this creature before?"). The creature used in the question in both conditions was either a creature presented in the preceding block or a novel creature matched in complexity. In the Math condition, the participants were asked a simple arithmetic question (“What is 5 + 7?”) in a multiple-choice format.

At the end of the eight blocks, participants were asked to rate the similarity between pairs of creatures and complexity of creatures they encountered on a 7-point Likert Scale. We used responses to these questions to make sure our complexity manipulation was successful, and check that we successfully controlled for similarity between backgrounds and deviants across complexity level

Participants We recruited 449 participants (Age M = 30.83; SD = 17.44) on Prolific. They were randomly assigned to one of the three conditions of the experiment (Curiosity: N = 156; Memory: N = 137; Math: N = 156). Participants were excluded if they showed irregular reaction times or their responses in the filler tasks indicates low engagement with the experiment. All exclusion criteria were pre-registered. The final sample included 380 participants.

Results

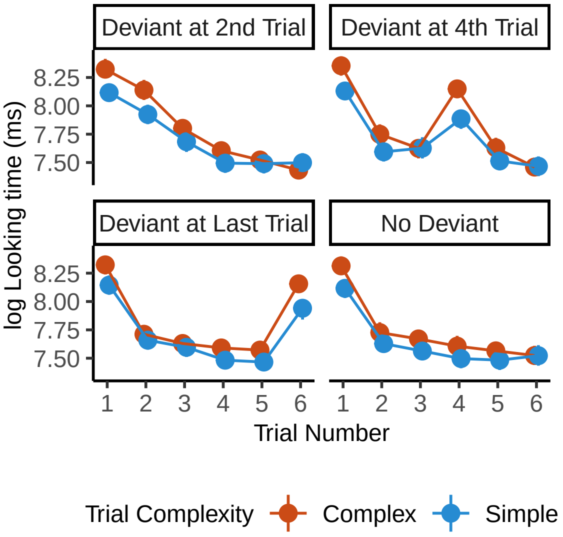

Figure 3: Results of behavioral experiment.

The sample size, methods, and main analyses were all pre-registered and are available at [LINK]. Data and analysis scripts are available at [LINK].

Manipulation Check The complex animated creatures were rated as more perceptually complex (M = ; SD =) than the simple animated creatures (M = ; SD =). Pairs of background creature and deviant creature were rated as moderately dissimilar to one another (M = ; SD =).

Evaluating the Paradigm Three criteria were selected to evaluate whether the paradigms successfully captured the characteristic looking time patterns observed in infant literature: habituation (the decrease in looking time for a stimulus with repeated presentations), dishabituation (the increase in looking time to a new stimulus after habituated to one stimulus), and complexity effect (longer looking time for perceptually more complex stimuli). The visualization of our results suggests that we reproduce the phenomena qualitatively (Fig. 3).To evaluate the phenomenon quantitatively, we ran a linear mixed effects model with maximal random effect structure. The predictor included in the model are a three-way interaction term between the trial number on an exponential function, the type of trial (background vs. deviant) and the complexity of the stimuli (simple and complex). The model failed to converge, so we pruned the model following the pre-registered procedure. The final model included subject as a random intercept. All predictors except for the three-way interaction are significant from the model (all $p < .001$). This provides a quantitative confirmation that our paradigm successfully captured the key looking time patterns.

Model comparison

To evaluate whether our models can provide sufficient explanation to the behavioral results, we designed a model experiment to represent the behavioral experiments. Then, we searched for the best set of parameters that yielded the highest Pearson’s correlation between the model results and behavioral results. We then compared the model fits within each model’s different linking hypotheses.

Model experiment

To model the behavioral experiment, we first represented the stimuli as a vector of logical values indicating the presence and absence of a feature. All stimuli vectors are length 6, with the complex stimuli represented as having three `TRUE` and simple stimuli represented as having one `TRUE`, The rest of the elements are `FALSE`. Individual stimuli are then assembled into sequences to reflect the stimuli sequences in the behavioral experiment. For a particular sequence, we constructed the deviant stimulus based on the background stimulus to make sure that they were always maximally different and had the same number of features present.

For Model 1, since it’s behavior is non-probabilistic, we presented the model with each of the four sequences once and derived the information theoretic measurements. For model 2, we ran each sequence 500 times to obtain a reasonably precise estimate on the model’s behaviors.

Parameter estimation

We performed an iterative grid search in parameter space for each linking hypothesis. We a priori constrained our parameter space on the prior beta distribution to have shape parameters that $\alpha_\theta > \beta_\theta$, which describe the prior beliefs as “more likely to see the absence of a feature than the presence of a feature.” For model 2, we searched for the priors over the concept (θ), the noise parameter that decides how likely a feature would be misperceived (ϵ), and the constant EIG from the world ($EIG(world)$). The prior over the noise parameter was fixed for all searches ($\alpha_\epsilon = 1; \beta_\epsilon = 10$). In model 2, different parameters were selected to obtain the best fit to the behavioral data (EIG: $\alpha_\theta = 1$, $\beta_\theta = 4$, $\epsilon = 0.065$, $EIG(world)$

= 0.01; KL: $\alpha_\theta = 1$, $\beta_\theta = 5$, $\varepsilon = 0.055$, $EIG(world) = 0.006$; Surprisal: $\alpha_\theta = 1$, $\beta_\theta = 3$, $\varepsilon = 0.07$, $EIG(world) = 8$).

Results

Discussion

General discussion

References

10 Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: Drift-diffusion model is equivalent to a bayesian model. *Frontiers in Human Neuroscience*, *8*, 102.

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS Computational Biology*, *17*(3), e1008863.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Haber, N., Mrowca, D., Fei-Fei, L., & Yamins, D. L. (2018). Learning to play with intrinsically-motivated self-aware agents. *arXiv Preprint arXiv:1802.07442*.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*.

Hunter, M. A., Ames, E. W., & Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology*, *19*(3), 338.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399.

Liquin, E. G., Callaway, F., & Lombrozo, T. (2021). Developmental change in what elicits curiosity. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, *11*(2), 265–286.

Poli, F., Serino, G., Mars, R., & Hunnius, S. (2020). Infants tailor their attention to maximize learning. *Science Advances*, *6*(39), eabb5053.

Sim, Z. L., & Xu, F. (2019). Another look at looking time: Surprise as rational statistical inference. *Topics in Cognitive Science*, *11*(1), 154–163.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, *22*(4), 325–336.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. *Advances in Neural Information Processing Systems*, 59–68.

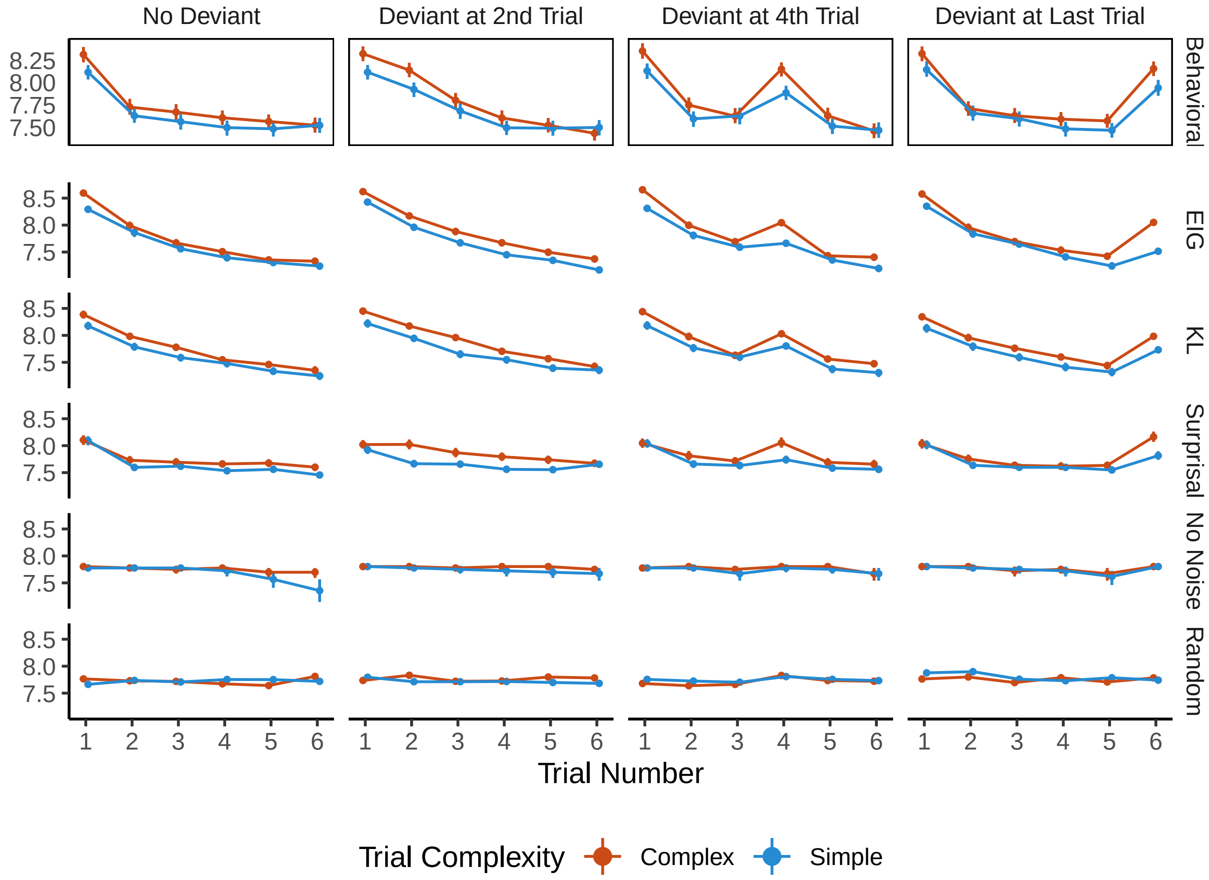

Figure 4: Continuous time model using different linking hypotheses provide qualitatively indistinguishable fits to the behavioral data. All model results are log-transformed and adjusted to be at the same scale and intercepts as the log-transformed behavioral data. The solid lines represent human data, and the dotted lines represent the model's results. Red lines indicated results for complex stimuli, and blue lines indicated results for simple stimuli.