A BAYESIAN RACE MODEL FOR RESPONSE TIMES UNDER CYCLIC STIMULUS DISCRIMINABILITY[1]

By Deborah Kunkel[*], Kevin Potter[†], Peter F. Craigmile[*,2],
Mario Peruggia[*] and Trisha Van Zandt[*]

Ohio State University[*] and University of Massachusetts[†]

Response time (RT) data from psychology experiments are often used to validate theories of how the brain processes information and how long it takes a person to make a decision. When an RT results from a task involving two or more possible responses, the cognitive process that determines the RT may be modeled as the first-passage time of underlying competing (racing) processes with each process describing accumulation of information in favor of one of the responses. In one popular model the racers are assumed to be Gaussian diffusions. Their first-passage times are inverse Gaussian random variables and the resulting RT has a min-inverse Gaussian distribution. The RT data analyzed in this paper were collected in an experiment requiring people to perform a two-choice task in response to a regularly repeating sequence of stimuli. Starting from a min-inverse Gaussian likelihood for the RTs we build a Bayesian hierarchy for the rates and thresholds of the racing diffusions. The analysis allows us to characterize patterns in a person's sequence of responses on the basis of features of the person's diffusion rates (the "footprint" of the stimuli) and a person's gradual changes in speed as trends in the diffusion thresholds. Last, we propose that a small fraction of RTs arise from distinct, noncognitive processes that are included as components of a mixture model. In the absence of sharp prior information, the inclusion of these mixture components is accomplished via a two-stage, empirical Bayes approach. The resulting framework may be generalized readily to RTs collected under a variety of experimental designs.

1. Introduction. Human performance data, measurements of the speed and accuracy with which tasks can be performed, is the foundation of most work in experimental psychology. Theories about behavior and how the brain processes information are evaluated on the basis of how human performance data change over different experimental conditions. Response times (RTs) in particular have been the foundation of much work in cognitive psychology for well over 100 years [Luce (1986)].

RTs in a typical sample are not independent and identically distributed (i.i.d.). Not only does the data generating mechanism change as the person becomes more

experienced with or tires from the task, but there are also short-term dependencies over consecutive responses arising from repetitions of stimuli or responses, making mistakes or other factors. Furthermore, RT data are not "clean." That is, a sample of RTs measured from an individual performing a task contains not only observations generated by the cognitive mechanism that is the focus of research but also "contaminant" observations resulting from inadvertent slips, momentary inattention and so forth. Such observations are more likely to be very fast or very slow and are thought to contribute to the heavy tails typically seen in the marginal distributions of RT sequences. Theoretically-motivated models are inappropriate for these contaminated trials. The inclusion of the more extreme observations may ultimately distort estimation of parameters of the true cognitive model, as discussed by Ratcliff (1993) and Whelan (2008).

Popular strategies for mitigating the influence of contaminants include transforming the data or, more commonly, discarding the most extreme RTs. Alternative approaches have avoided removing data by modeling RTs as a mixture of a cognitive process and two additional processes that account for the short and long responses. Vandekerckhove and Tuerlinckx (2007) used such a model with uniform distributions on the mixture components, and, more recently, Kim et al. (2017) modeled the contaminants using log-normal distributions, where informative priors on these distributions were used to promote a stochastic ordering of the mixture components.

While many researchers make the assumption of i.i.d., uncontaminated observations out of computational convenience, Craigmile, Peruggia and Van Zandt (2010) proposed a modeling framework that deliberately incorporated these well- i.i.d. characteristics of the data. Their purpose was to obtain better estimates of the parameters of the data-generating mechanism by using a more complete model that included mechanisms by which contaminant observations, changes due to experimental conditions, and trial-by-trial dependences might arise.

In the descriptive model proposed by Craigmile, Peruggia and Van Zandt (2010), log RTs were generated by a simple autoregressive process of order 1 [AR(1)]. In this model the distribution of a log RT on any given trial was a function of the RT on the preceding trial. In addition there was some probability that the RT would be delayed or sped up by some amount, the amount determined by an exponential random variable that was either added to or subtracted from the log RT generated by the AR(1) process. Long-term trends over trials were described using a wavelet regression added to the log RT. This model was embedded in a Bayesian hierarchy and fit to data from a simple detection task published by Wagenmakers, Farrell and Ratcliff (2004), and model diagnostics suggested that the model accounted for the data well.

The research in the present paper builds on the modeling framework presented by Craigmile, Peruggia and Van Zandt (2010) in three major ways. First, we explore a theoretically motivated data-generating mechanism with more explanatory power than the AR(1) process originally proposed. This mechanism supposes that

simple decisions arise from a race between two diffusion processes and that RTs are determined by the first-passage time of the process that reaches a fixed threshold first. This kind of mechanism has seen wide success in explaining simple decision processes in cognitive psychology [Logan et al. (2014), Usher and McClelland (2001)] and has received some support from observations of brain activity in studies involving both human and nonhuman primates [Heitz and Schall (2012)].

Second, we test the utility of this kind of modeling framework by using it to recover "signals" in the stream of measured RTs induced by regularities in the sequence of stimuli presented to the students who participated in the experiment. We manipulated the response-relevant visual features of the stimulus sequence so that changes in stimulus discriminability induced by relative color intensity and distance from the center of the computer monitor varied cyclically over trials. A major test of the modeling framework is to try and recover the "footprints" of these signals in the data free from distortion due to outlying contaminant RTs.

Third, we accommodate contaminant RTs using a mixture model. In the absence of sharp prior information about the processes that generate these observations, we use a two-stage procedure that begins with a preliminary model that includes only a cognitive component. We then identify ill-fitting points that are likely to have been generated from the noncognitive processes. We use these points and an empirical Bayes approach to build data-dependent priors for the parameters of the mixture components.

In Section 2 we present a brief description of the experiment we conducted. The data from this experiment were used to develop and test the hierarchical model that we present in Section 3. We summarize the fits of the model in Section 4 and discuss model validation in Section 5. Finally, in Section 6 we present our findings and outline directions for possible extensions. The Supplementary Material [Kunkel et al. (2019)] provides additional detail on the experimental procedure, displays the RT sequences for all students and reports the results of the analysis for all students. The Supplementary Material also includes a demonstration of the ability of our estimation procedure to recover the parameter values used to generate simulated data for a prototypical experimental participant, a comparison of the performance of our *theoretically-motivated* modeling framework with that of a *descriptive* generalized additive model and an evaluation of the predictive performance of our approach.

2. Description of the experiment. Data were collected from 22 Ohio State University (OSU) undergraduates recruited from introductory psychology courses through the Research Experience Program at OSU. The experiment asked these students to quickly judge whether a small square was presented to the left or right of the center of a computer monitor. The difficulty of the discrimination was varied by: (i) presenting the square at varying distances from the center (3, 10, 30 or 80 pixels from the screen center) and (ii) varying the contrast between the square
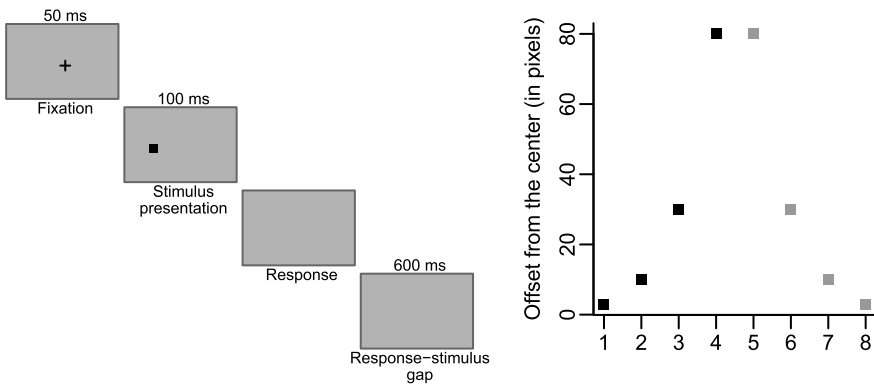
FIG. 1.   *Experimental procedure. The left panel shows the structure of a single trial for the task. The duration for each frame is shown above each rectangle, while the frame's role in the trial is stated below. The right panel shows the sequence of horizontal offsets and relative color intensities used for the target stimulus over one segment of trials.*

and the background (gray levels of 0 [black] or 115 [gray] against a background gray level of 127 [darker gray]).

Students completed a series of 552 trials with no rest breaks. The left panel of Figure 1 presents the structure of a single trial from the task. The right panel of Figure 1 shows the sequence of stimuli over trials. The stimulus appeared randomly to the left or right of center, while the sequence of eight stimuli (varying in distance and color intensity) was repeated over the 552 trials, resulting in a pattern of stimulus changes with periods 8.

We discarded the first 40 trials as a "practice" period. The resulting series of 512 log RTs are shown in Figure 2 for Students 3, 5, 6, 10, 19 and 21. For ease of presentation, we focus our figures and discussion on these students. We chose Students 3, 5, 10, 19 and 21 because our analysis revealed that they were representative of several interesting features also found in other study participants and Student 6 because the analysis revealed this student to have anomalous, idiosyncratic features not shared by any of the other students in the study. The Supplementary Material contains the entire data set and a detailed description of the experimental design.

3. Hierarchical Bayesian min-inverse Gaussian race model. Modeling the accumulation of evidence over time as a stochastic process has been very successful in providing explanations for a wide range of behaviors in tasks as diverse as recognition memory, signal detection, pattern matching and economic decisions [Caplin and Martin (2016), Ratcliff and McKoon (2007), Van Zandt, Colonius and Proctor (2000)]. In particular, models based on Gaussian processes have been fit to human performance data to explain how decision speed and accuracy trade off against each other and have also formed the basis of models that connect behavior
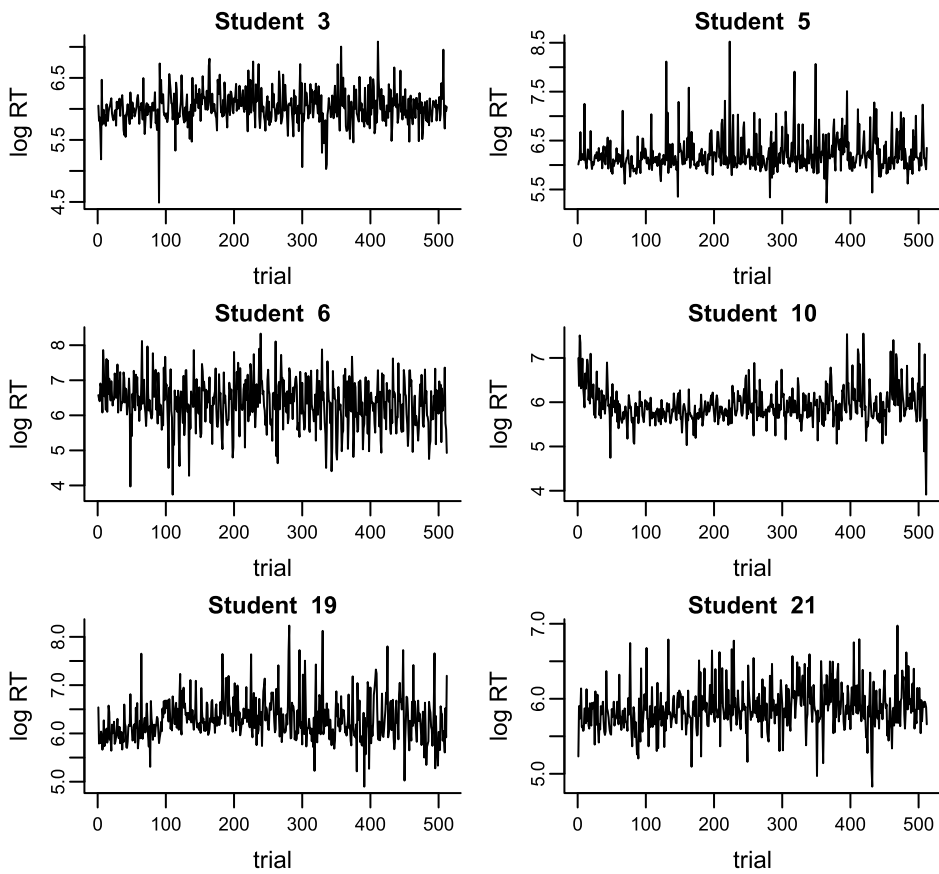
FIG. 2. *Observed* log *RTs for Students* 3, 5, 6, 10, 19, *and* 21.

to specific neural events [Nelson, Murthy and Schall (2016)]. While other kinds of stochastic processes have been used to describe human performance data [Brown and Heathcote (2008), Van Zandt, Colonius and Proctor (2000)], it is generally agreed that those based on continuous Gaussian processes provide the best fits to the data [Ratcliff, Smith and McKoon (2015)].

A class of these models—"race" models—casts the decision process as a race between independent accumulators with each one representing a possible response to a given problem. The first accumulator to reach a fixed threshold determines the response to be made, and its first-passage time determines the RT. With two possible responses in this experiment, this model states that the distribution of RTs, without consideration of contaminants, sequential effects or long-term trends, should be distributed as the minimum of two first-passage times, one corresponding to each of the possible responses to the stimulus. Further, we specify, consistent with prior research [Logan et al. (2014)], that the accumulation process be

described as Brownian motion, so that the first-passage time distributions both follow inverse Gaussian distributions, and the RT is determined by the minimum of the two, the min-inverse Gaussian race model.

3.1. *Min-inverse Gaussian race model.*   The min-inverse Gaussian race model describes observed RTs as a minimum of two inverse Gaussian random variables. In the context of our experiment, this model arises if we assume that two racing accumulators, corresponding to "right" or "left" responses, are each described as Brownian motion processes with drifts v^s, where the superscript $s = l$ or r refers to the left (l) or right (r) response process. The first-passage time of one such process at a threshold α^s is distributed as an inverse Gaussian variable with probability density function

$$(3.1) \qquad f(y|\alpha^s, v^s) = \left(\frac{\alpha^s}{2\pi y^3}\right)^{1/2} \exp\left(\frac{-(v^s)^2(y - \alpha^s/v^s)^2}{2y\alpha^s}\right), \qquad y > 0.$$

Under this distribution the first-passage time has a mean of α^s/v^s and variance of $(\alpha^s)^2/(v^s)^3$. This distribution accommodates noisy accumulation of information in favor of a particular response and its parameters are directly interpretable as mechanisms in the race model. The diffusion rate v^s describes how quickly information is accumulated, whereas the threshold α^s characterizes the amount of evidence necessary to conclude that the response is the correct one [Logan et al. (2014)]. Slower first-passage times may result from a decrease in the drift v^s, an increase in the threshold α^s or both.

In this model responses are generated from the racing "left" and "right" processes, whose first-passage times at trial t are inverse Gaussian random variables which may have different rates and thresholds. We observe an RT Y_t, which is the smaller of the two first-passage times, and S_t, an indicator of which process won at trial t which equals l if the student's response was "left" and r if the response was "right."

At trial t, the joint density of (Y_t, S_t) is

$$(3.2) \qquad f(y_t, s_t) = \begin{cases} f_l(y_t)(1 - F_r(y_t)) & \text{if } s_t = l, \\ f_r(y_t)(1 - F_l(y_t)) & \text{if } s_t = r, \end{cases}$$

where $f_l = f(y_t \mid \alpha^l, v^l)$ and $f_r = f(y_t \mid \alpha^r, v^r)$ are the inverse Gaussian densities for the left and right processes respectively, and F_l and F_r are the corresponding cumulative distribution functions. We will refer to (3.2) as the "min-inverse Gaussian" distribution.

3.2. *Hierarchical model for racing processes.*   For Student j, $j = 1, \ldots, 21$, and trial t, $t = 1, \ldots, 512$, we observe the RTs and responses (Y_{jt}, S_{jt}).

The pairs (Y_{jt}, S_{jt}) are distributed as min-inverse Gaussian random variables and are conditionally independent given the thresholds and rates of the left and

right processes. We write these parameters for Student j at trial t as $(\alpha^s_{jtu}, v^s_{jtuc})$, where the subscript $u = l$ (left) or r (right) refers to the side at which the stimulus was presented at trial t, and the color intensity is either $c = b$ (black) or g (gray). We assume that the two processes are symmetric, so that

$$(3.3) \qquad (\alpha^l_{jtl}, v^l_{jtlc}) = (\alpha^r_{jtr}, v^r_{jtrc}) \quad \text{and} \quad (\alpha^r_{jtl}, v^r_{jtlc}) = (\alpha^l_{jtr}, v^l_{jtrc}),$$

for all j, t, c; that is, the rate and threshold of process s depend only on the side on which the stimulus was presented.

When referring to the process s that controls the response that is consistent with the side on which the stimulus u is presented, that is, when $u = s$, we will use the term "correct process." When $u \neq s$, we will use the term "incorrect process."

We model the log rate for Student j, trial t, side u, color intensity c and process s as a baseline rate μ^s_{juc} to which we add a conditionally independent term for the correct process:

$$(3.4) \qquad\qquad \log(v^s_{jtuc}) = \begin{cases} \mu^s_{juc} + b^s_{ju}(t) & \text{if } s = u, \\ \mu^s_{juc} & \text{if } s \neq u. \end{cases}$$

The term $b^s_{ju}(t)$ is a periodic signal that is a function of trial t and accounts for the changing difficulty of the task due to the distance of the presented stimulus from the center.

We first present a hierarchical model for the baseline rates μ^s_{juc}. We next present the model for the signal $b^s_{ju}(t)$ which incorporates the effects of stimulus distance from center. We will then present a semiparametric model for the thresholds $\log(\alpha^s_{jtu})$. The hierarchical structure on the rates permits dependence between rates for trials whose stimuli are similar in color intensity or location. The hierarchy also permits dependence between thresholds across trials to account for sequential effects.

Model for baseline rates μ^s_{juc}. All trials for Student j, stimulus side u and color intensity c share a common baseline rate μ^s_{juc} which may be different for the correct and incorrect processes. We assume that for process s and side u, the baseline rates μ^s_{jub} and μ^s_{jug} are conditionally independent and distributed as

$$\mu^s_{jub} \sim N\big(\mu^s_{ju} + \gamma^s_u, (\tau^s_{\mu ju})^2\big) \quad \text{and} \quad \mu^s_{jug} \sim N\big(\mu^s_{ju} - \gamma^s_u, (\tau^s_{\mu ju})^2\big),$$

where $N(a, b)$ denotes a normal distribution with mean a and variance b. Here $2\gamma^s_u$ represents an additive effect of the darker intensity on the log rate for process s and side u, and μ^s_{ju} is the student-specific mean rate for process s and side u. The difference $\mu^s_{jub} - \mu^s_{jug}$ captures the student-specific mean effect of color intensity change for the correct process when $s = u$ and for the incorrect process when $s \neq u$.

We give each effect of color intensity γ^s_u a normal prior with mean γ and variance τ^2_γ. The baseline rates μ^s_{ju} have normal priors with means $\mu_j + \delta$ ($s = u$) and

TABLE 1
Hyperpriors on the model parameters for μ^s_{juc}

$(\tau^s_{\mu j u})^{-2}$	$\tau^{-2}_{\mu j}$	τ^{-2}_{μ}	τ^{-2}_{γ}	γ	δ	μ
$G(8, 2)$	$G(8, 2)$	$G(1, 1)$	$G(1, 1)$	$N(0, 20)$	$N(0, 20)$	$N(0, 20)$

$\mu_j - \delta$ ($s \neq u$) and variance $\tau^2_{\mu j}$. The student-specific precisions $(\tau^s_{\mu j u})^{-2}$ and $\tau^{-2}_{\mu j}$ are i.i.d. $G(8, 2)$ for all students, where $G(a, b)$ is a gamma distribution with mean a/b and variance a/b^2. These distributions put high probability on higher precisions and make it unlikely, a priori, for the baseline rates of the two processes to be vastly different. The means μ_j are normal with mean μ and common variance τ^2_{μ}. The prior distribution of the remaining hyperparameters are given in Table 1, where conditional independence is assumed throughout. The hyperparameters for τ^{-2}_{μ}, τ^{-2}_{γ}, γ, δ and μ were chosen to be somewhat diffuse to reflect our prior ignorance about these parameters. In future studies more informative choices could be made using, in part, the additional information gathered from this study. While, formally, the posterior distribution from the current study should become our future prior, we do not fully endorse such an approach due to the complexities of the model and the data. Rather, we prefer to recommend a more cautious approach in which the distributions of the hyperparameters are kept fairly diffuse.

Model for effects of stimulus distance $b^s_{ju}(t)$. As we explained in Section 2, students in this study were presented with a sequence of stimulus distances and intensities in repeating cycles of eight trials, with randomly chosen left or right side placement.

When the stimulus was presented on side u, we expect the diffusion rate for the correct process $s = u$ to increase with increasing distance, reflecting the greater ease with which the left/right discrimination could be made and the corresponding faster RTs. We assume that distance changes on side u do not influence the rate of the incorrect process $s \neq u$.

To account for possible periodicities, we defined $b^s_{ju}(t)$ in (3.4) as a sinusoidal signal using a harmonic regression at a set of four prespecified frequencies, $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$, where $f_i = 1/2^i$ for each i. This harmonic regression model for Student j, side u and process s is given by

$$(3.5) \quad b^s_{ju}(t) = \sum_{i=1}^{4} (\beta^s_{1iju} \cos(2\pi f_i t) + \beta^s_{2iju} \sin(2\pi f_i t)), \qquad t = 1, \ldots, 512.$$

The amplitude A^s_{iju} and phase ω^s_{iju} at the ith frequency are determined by the coefficients β^s_{1jiu} and β^s_{2iju} as follows:

$$A^s_{iju} = \sqrt{(\beta^s_{1iju})^2 + (\beta^s_{2iju})^2}, \qquad A^s_{iju} > 0;$$

$$\omega_{iju}^s = \tan^{-1}\left(\frac{\beta_{2iju}^s}{\beta_{1iju}^s}\right), \qquad \omega_{iju}^s \in [0, 2\pi).$$

The amplitude A_{3ju}^s at frequency $f_3 = 1/8$ reflects the sensitivity of the rate for process s to the changes in stimulus distance. The remaining amplitudes A_{iju}^s reflect the magnitude of predictable variation in the rates that cannot be easily attributed to the stimuli.

Let $\boldsymbol{\beta}_{1ju}^s$ denote the vector whose elements are the coefficients β_{1iju}^s for frequencies $i = 1, \ldots, 4$, and let $\boldsymbol{\beta}_{2ju}^s$ be defined analogously. Because the sine and cosine functions are orthogonal, we assume that $\boldsymbol{\beta}_{1ju}^s$ and $\boldsymbol{\beta}_{2ju}^s$ are conditionally independent and normally distributed with mean vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively and covariance matrix $\tau_\beta^2 \boldsymbol{D}$ for $j = 1, \ldots, 21$. Here $\boldsymbol{D}$ is a 4×4 diagonal matrix whose ith diagonal entry is $1/(2f_i)$, where f_i is the ith frequency in $\mathcal{F}$.

This specification for the covariance matrix $\boldsymbol{D}$ assigns larger prior variance to coefficients associated with lower frequencies (longer wavelengths). We further assign independent $N_4(\mathbf{0}, 6\boldsymbol{D})$ distributions to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, where $\mathbf{0}$ is a vector of 0s. We give the precision parameter τ_β^{-2} a G(1, 1) prior.

Model for response thresholds α_{jtu}^s. In addition to regularly repeating changes arising from experimental conditions, the series of RTs also exhibit local trends such as unusually slow responses at the beginning or end of the experiment, or periods of consecutive fast responses. Several of the students' log RT series in Figure 2 show such trends, for example, Students 3 and 19 appear to answer quickly at the beginning of the experiment and slow down over time, Student 6 appears to speed up over time, and Student 10 has unusually slow responses in the first 50 trials. These behaviors may be due to factors such as fatigue, periods of delayed response and learning; it is hard to attribute them to the systematic changes in the experimental design.

Within the context of our model, these features cannot be explained by the log-linear expansion for $\log v_{jtuc}^s$. They can, however, be explained by incorporating smooth trends in the thresholds α_{jtu}^s. Because the mean and variance of the inverse Gaussian distribution increase with α_{jtu}^s and $(\alpha_{jtu}^s)^2$ respectively, RTs and their variability will tend to be large during periods when α_{jtu}^s is elevated. We performed an exploratory data analysis calculating the moving averages of the means and variances of the RT sequences that supported the need for these trends.

We use $\boldsymbol{\alpha}_{ju}^s$ to denote the vector containing α_{jtu}^s, $t = 1, \ldots, 512$, the thresholds for Student j, process s and stimulus u at trial t. We model trends in $\log(\alpha_{jtu}^s)$ via a cubic spline regression on trial number using a B-spline basis. This basis uses nine basis functions that span the same space as $\{1, t, t^2, t^3, (t - \kappa_1)^3, \ldots, (t - \kappa_5)^3\}$, where $\kappa = \{\kappa_1, \ldots, \kappa_5\}$ are evenly spaced knots [Ruppert, Wand and Carroll (2003)]. The log thresholds can be represented as

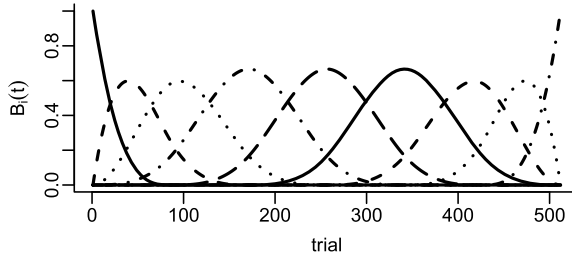$$(3.6) \qquad \log(\alpha_{jtu}^s) = \sum_{i=1}^{9} \phi_{iju}^s B_i(t),$$

FIG. 3. *Basis functions used to model* $\log(\boldsymbol{\alpha}_{jtu}^s)$.

where $B_i(t)$ is the ith basis function evaluated at trial t. We constructed this basis using the R "splines" library with a default choice of knots at the 17th, 33rd, 50th, 67th and 83rd quantiles of the trial numbers; $\kappa = (86.2, 171.3, 256.5, 341.7, 426.8)$. Figure 3 shows these functions for the sequence of 512 trials. We chose to use five knots because this choice of basis allows local flexibility in the threshold while keeping the number of parameters modest. A sensitivity analysis using several different choices of knots showed that the predictive properties of the cognitive model was not sensitive to this choice.

We denote by $\boldsymbol{\phi}_{ij} = (\phi_{ijl}^l, \phi_{ijl}^r)^T = (\phi_{ijr}^r, \phi_{ijr}^l)^T$ the vector of coefficients associated with $B_i(t)$ for Student j's correct and incorrect processes. For each j we assume that the prior for $\boldsymbol{\phi}_{ij}$, $i = 1, \ldots, 9$, is a bivariate normal distribution with mean $\mu_{\phi j} \mathbf{1}_2$ and covariance matrix $\tau_\phi^2((1 - \rho_\phi)\boldsymbol{I}_2 + \rho_\phi \boldsymbol{J}_2)$, where $\mathbf{1}_2$ is a column vector of ones, $\boldsymbol{I}_2$ is the 2×2 identity matrix, and $\boldsymbol{J}_2$ is a 2×2 matrix of ones. This defines an exchangeable correlation structure where, for Student j, ϕ_{iju}^l and ϕ_{iju}^r have correlation ρ_ϕ for a fixed basis function i, and ϕ_{iju}^l and $\phi_{i'ju}^r$ are uncorrelated for $i \neq i'$. The student-specific mean threshold is captured by $\mu_{\phi j}$, whose prior is a normal distribution with mean μ_ϕ and variance τ_ϕ^2, where μ_ϕ and $\tau_{\mu_\phi}^{-2}$ have $N(0, 20)$ and $G(1, 1)$ distributions respectively.

3.3. *Mixture likelihood to account for contaminant RTs.*   As discussed in the Introduction, RTs are "contaminated" by unusually long and short responses that are generated by processes unrelated to the cognitive process of interest. We accommodate these contaminants with a mixture model that generates observations from the min-inverse Gaussian cognitive model with high probability, and from two other processes, G and H, with small probability (η_G and η_H respectively). The process G describes the faster ("subcognitive") RTs, and H generates the unusually long ("supracognitive") RTs. The full mixture likelihood for (Y_{jt}, S_{jt}) is

$$(3.7) \quad P(Y_{jt} \leq y, S_{jt} = s) = (1 - \eta_{Gj} - \eta_{Hj})F(y, s) + \eta_{Gj}G(y) + \eta_{Hj}H(y),$$

where F is the min-inverse Gaussian distribution in (3.2) and G and H denote the cumulative distribution functions of the log-normal distribution. The parameters

of G are fixed and determined by prior information, and the parameters of H are student-specific. As we next explain, the prior distributions for the parameters of H are chosen using prior knowledge and diagnostic information from the data.

3.3.1. *Models for G and H*.   We take G, the model for subcognitive RTs, to be a log-normal distribution such that the log RTs have a mean of log(150) and standard deviation of 0.25 denoted by log-normal(log(150), 0.25). This choice is motivated by existing work on "fast outliers," which suggests that RTs below 100 ms are almost certainly not generated from a cognitive process, and those lower than 200 ms are also likely to be guesses [Ratcliff (1993), Whelan (2008), Baayen and Milin (2010), Kim et al. (2017)]. Our choice for G places about 98% of its mass below 250 ms.

The supracognitive component H is a log-normal(a_j, b_j) distribution for Student j, where the student-specific parameters a_j and b_j, $j = 1, \ldots, 21$, are conditionally independent and drawn from priors P_a and P_b respectively. We take P_a to be a log-normal(μ_a, σ_a) distribution and P_b to be a log-normal(μ_b, σ_b) distribution.

Choosing the appropriate values for the hyperparameters μ_a, σ_a, μ_b and σ_b is not straightforward. Little prior information is available about the appropriate location shift for H, and the features of H may vary across individuals. Fully noninformative priors cannot be used, however, because the model is weakly identified unless we use the prior to distinguish H from the cognitive model. To address this problem, we take a two-stage, empirical Bayes approach and use a reweighted version of the observed data to estimate μ_a, σ_a, μ_b and σ_b.

Our approach is as follows: for Student j, we assign a weight $0 \leq w_{jt} \leq 1$ to each observed RT at trial t that reflects our empirical belief that this observation may be generated from H. We then calculate weighted maximum likelihood (ML) estimates of the log-normal location and shape parameters using the weighted likelihood $\prod_{t=1}^{512} f(y_{jt}|a_j, b_j)^{w_{jt}}$. We denote these ML estimates by $\widehat{a}_j$ and $\widehat{b}_j$. Last, we set μ_a and σ_a to equal the log-normal ML estimates calculated from the $\widehat{a}_j$. Similarly, μ_b and σ_b are derived as ML estimates based on the $\widehat{b}_j$.

In the paragraphs below, we first describe how we estimated the likelihood weights w_{ij}. We then describe the model for the mixture probabilities η_G and η_H.

Method for calculating likelihood weights w_{jt}. To calculate the weights w_{jt}, we use three principles from the RT literature: first, very slow RTs are likely to have been generated from H; second, observations that fit poorly in the cognitive model are likely to have been generated from H; third, we do not know with certainty whether any single observation was or was not generated from H. These principles are consistent with discussions found in, among others, Ratcliff (1993), Whelan (2008) and Baayen and Milin (2010) in the context of deciding which observations should be trimmed from an RT dataset. Baayen and Milin (2010) in particular focus on excluding those extreme RTs that seem to contribute to a lack of fit in the
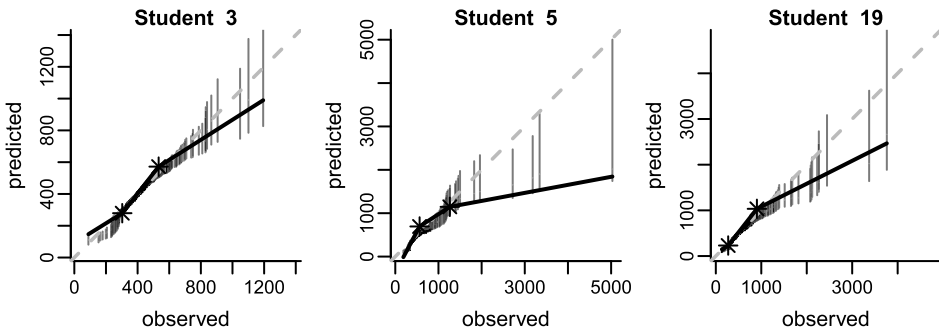
FIG. 4.   *Q–Q plots for the preliminary fit of the cognitive model only. The dashed line is the identity line, and the solid line is the piecewise linear fit using equation* (3.8). *The asterisks denote the fitted values at* $y_{j(q_1)}$ *and* $y_{j(q_2)}$.

cognitive model; we build on this principle by using model diagnostics to quantify the lack of fit for each observation and using this information to calculate w_{jt}.

To this end we fit a restricted model in which $\eta_{Gj} = \eta_{Hj} = 0$; only the min-inverse Gaussian component of the mixture was fit to the RTs. We evaluated the fit of this model by comparing the observed RTs with samples from its posterior predictive distribution. For most students the slowest RTs were much slower than those predicted by the min-inverse Gaussian model, as exhibited by poor coverage of posterior intervals and by quantile–quantile (Q–Q) plots that show misfit of the tails of the distribution. These Q–Q plots, shown for Students 3, 5 and 19 in Figure 4, exhibit model fits that deteriorate sharply after a point in the upper tail, as indicated by an "elbow" in the plot. The points at or above this elbow should contribute the most information about the parameters of H.

To isolate this subset of the data, we used a piecewise linear model of predicted quantiles on observed quantiles. Let $y_{j(t)}$ be the $t/512$th sample quantile of the observed data for Student j and use $\tilde{y}^b_{j(t)}, b = 1, \ldots, 300$, to denote 300 samples from the posterior predictive distribution of that quantile. The following weighted least-squares regression of $\tilde{y}^b_{jt}$ on $y_{j(t)}$ allows three connected line segments to comprise the best-fit response function,

$$(3.8) \quad \begin{aligned} \tilde{y}^b_{j(t)} &= \beta_0 + \beta_1 y_{j(t)} + \beta_2 (y_{j(t)} - y_{j(q_1)}) + \beta_3 (y_{j(t)} - y_{j(q_2)}) + \varepsilon^b_{j(t)}, \\ \varepsilon^b_{j(t)} &\sim N(0, \sigma^2 v_{j(t)}), \qquad b = 1, \ldots, 300, \end{aligned}$$

where $y_{j(q_1)}$ and $y_{j(q_2)}$, the $q_1/512$ and $q_2/512$ sample quantiles of $\mathbf{y}_j$, are the knots at which the slope of the line is allowed to change. We set the least-squares weights $v_{j(t)}$ in this segmented regression model to be proportional to an estimate of the posterior variances of $\tilde{y}_{j(t)}$. For each student we identified the values of q_1 and q_2 that minimized the residual sum of squares and determined which observations fell above the upper "elbow" $y_{j(q_2)}$. These points provide the most information about H and should have large weights.
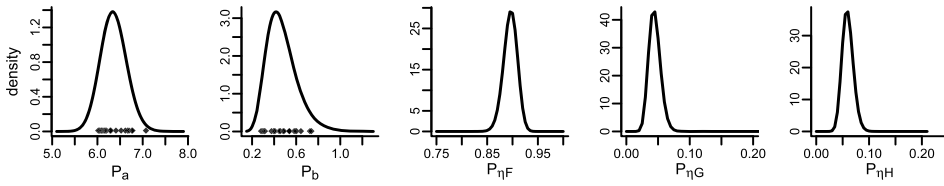
FIG. 5. *Densities of the prior distributions P_a and P_b, $P_{\eta,F}$, $P_{\eta,G}$ and $P_{\eta,H}$. The gray diamonds are $\widehat{a}_1, \ldots, \widehat{a}_{21}$ and $\widehat{b}_1, \ldots, \widehat{b}_{21}$ in the first and second panels respectively.*

The sampling weight w_{jt}, which determines each observation's weight in estimating the hyperparameters of P_a and P_b, was made to increase with the magnitude of the observation, $d_{jt} = y_{jt}/\max_h(y_{jh})$. Specifically, we let

$$(3.9) \qquad w_{jt} = \exp(c_1 d_{jt})/(c_2 + \exp(c_1 d_{jt})).$$

We computed the constants c_1 and c_2 by setting the desired weight for the $q_1/512$ and $q_2/512$ sample quantiles equal to 0.05 and 0.5 respectively and solving the resulting system of two equations. With this selection of the sampling weights, we use a weighted log-normal likelihood to calculate the weighted log-normal ML estimates $\widehat{a}_1, \ldots, \widehat{a}_{21}$ and $\widehat{b}_1, \ldots, \widehat{b}_{21}$. Finally, the parameters μ_a and σ_a were set equal to the ML estimates calculated from $\widehat{a}_1, \ldots, \widehat{a}_{21}$, and μ_b and σ_b were set equal to the ML estimates calculated using $\widehat{b}_1, \ldots, \widehat{b}_{21}$.

Many other reasonable methods could be used to calculate the sampling weights. The exact functional form of Expression (3.9) is somewhat ad hoc; however, the parameter recovery study reported in Section 3 of the Supplementary Material lends credibility to the method, because the proposed procedure led to increased accuracy in parameter estimation. Further, the sensitivity analysis presented in Section 5 suggests that, because the procedure is not very sensitive to changes of the hyperparameters μ_a, σ_a, μ_b and σ_b, the specifics governing the calculation of the weights are, within reason, unlikely to influence the conclusions of the analysis.

Priors on η_j. For P_η, the prior on $\boldsymbol{\eta}_j = (\eta_{Fj} = 1 - \eta_{Gj} - \eta_{Hj}, \eta_{Gj}, \eta_{Hj})$, we specified a Dirichlet distribution with hyperparameter $\boldsymbol{m} = (m_F, m_G, m_H)$ whose marginals we denote by $P_{\eta,F}$, $P_{\eta,G}$ and $P_{\eta,H}$. As with the hyperparameters of P_a and P_b, we required concentrated priors on these mixture proportions to impose identifiability between the flexible log-normal component and the min-inverse Gaussian distribution. Ratcliff (1993) has recommended that at least 85%–90% of observations be classified as arising from the cognitive model, in part to ensure that the long tails of the cognitive model are fully described. We chose $\boldsymbol{m} = (450, 22.5, 30)$ which gives a concentrated prior with expected proportions of $0.90, 0.04, 0.06$ of components F, G and H respectively. Figure 5 shows a graphical summary of the priors P_a, P_b, $P_{\eta,F}$, $P_{\eta,G}$ and $P_{\eta,H}$.

Two key aspects of our model specification are the theoretically motivated min-inverse Gaussian race model describing the cognitive RTs and the two-stage, data-informed procedure for specifying the distributions of the sub- and supracognitive RTs. To ensure that the information in the data is sufficient to learn about these two aspects of the model, we conducted a proof of concept experiment in which we analyzed simulated RTs for a prototypical student, with and without contaminant sub- and supracognitive RTs. The results of the experiment, reported in the Supplementary Material, confirm that, in the absence of contaminants, the parameters of the min-inverse Gaussian race model can be accurately recovered. We demonstrate that if contaminants are present, the min-inverse Gaussian race model alone cannot describe adequately the variability in the data, but, once the sub- and supracognitive mixture component are added and the two-stage specification procedure is applied, the parameter values used to simulate the data can again be recovered accurately.

4. Results. We fit the full mixture model using Markov chain Monte Carlo (MCMC) with a Gibbs sampling algorithm. We ran multiple chains for 350,000 iterations after burn-in and thinned the chains by keeping every 500th sample. We assessed convergence using trace plots and running multiple chains from different starting values. Here, we discuss our results, focusing on the six selected students.

Effects of stimulus changes. Underlying "signals" due to stimuli, if present, are captured in the posterior distribution of the log diffusion rates $\log(\nu)$. If a student's RTs respond predictably to changes in distance, we expect to see this reflected in the rate as a strong signal at the frequency 1/8. We similarly evaluate sensitivity to changes in the color intensity by looking for shifts in the posterior distribution of the log rate corresponding to the black and gray stimuli.

Figure 6 summarizes the posterior distributions of the log rates $\log(\nu^s_{jtuc})$ for the six selected students. The rates have been aligned so that a given rate index on the horizontal axis corresponds to the same experimental condition for all students, ordered as in the illustration in the right-hand panel of Figure 1. The plots on the left show the log rate for the correct process at trial t [i.e., $\log(\nu^l_{jtlc})$ or $\log(\nu^r_{jtrc})$] and those on the right show the log rates for the incorrect process [$\log(\nu^l_{jtrc})$ or $\log(\nu^r_{jtlc})$]. Except for Student 6, the patterns seen in this figure are generally representative of all the students who participated in the experiment.

We focus first on the correct process. In general the rate drops substantially when the stimulus shifts from black to gray on trial 5. This downward shift suggests that, although the stimulus at this setting is presented at the largest distance of 80, the discrimination is not easy because of the difficulty introduced by the switch to the lower color intensity. For many of the students, the rate is in fact slowest for this first trial after the switch.

There are individual differences in how students reacted to changes in distance. Some students (3, 19 and 21) showed a strong effect of decreasing distance when
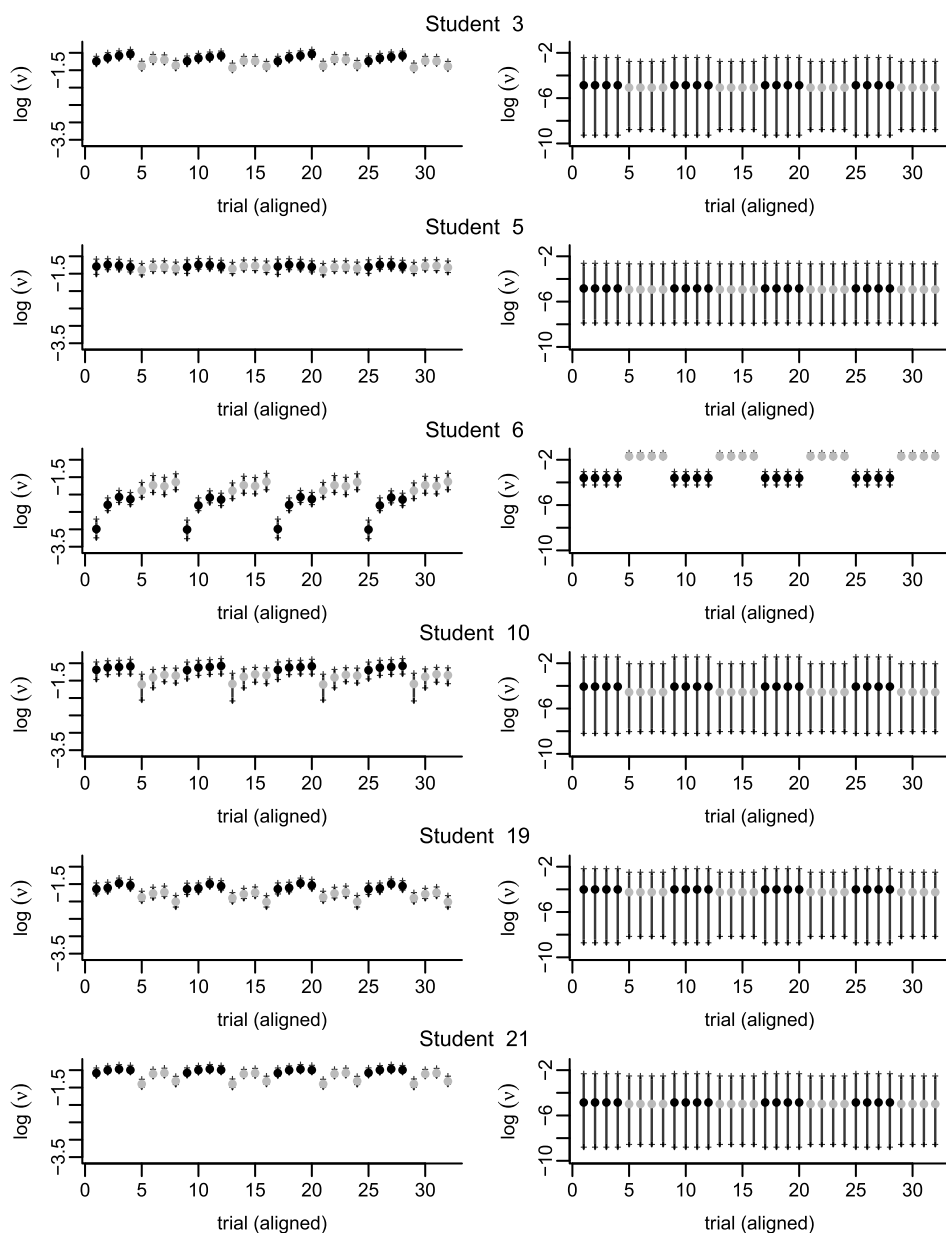
FIG. 6.   *Posterior means and* 90% *posterior intervals of log rates for selected students for stimuli presented in the order shown in Figure* 1. *The plots on the left show the log rate for the correct process* [$\log(v^l_{jtlc})$ *or* $\log(v^r_{jtrc})$]. *The plots on the right show the log rate for the incorrect process* [$\log(v^l_{jtrc})$ *or* $\log(v^r_{jtlc})$]. *The color intensity of the plotting symbol* (*black or gray*) *matches that of the stimulus.*
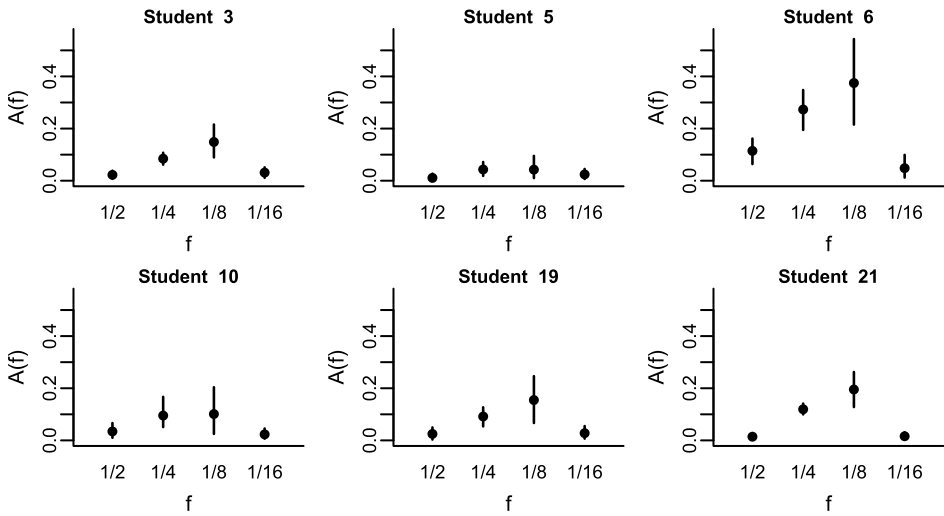
FIG. 7.   90% *posterior intervals of amplitudes* $b_j^s(t)$ *at each frequency for selected students.*

the stimulus was gray, with a large reduction in rate for a distance of 3. When the stimulus was black, many students were less sensitive to distance. Student 5's rate slowed slightly as distance decreased, while Student 3's rate increased with distance. The rates for Student 6, on the other hand, were smaller than those of the other students and changed perceptibly with distance when the stimulus was black but remained more stable and comparable to those of the other students when the stimulus was gray. Overall, the rates for the black stimuli were lower than those for the gray stimuli, a pattern that is unique to this student. For all students the changes in rates over each sequence of eight trials were nearly identical, indicating that the periodic component is weak at frequency 1/16 and stronger at higher frequencies.

For the incorrect process the right column of Figure 6 shows that the log rates tend to be smaller than for the other process and remain nearly constant across trials. Our model for these rates includes a shift for the change in stimulus color intensity. With one exception this shift is small for all students, indicating that the diffusion rate for the incorrect process is not very sensitive to any changes in the stimulus. For Student 6, however, the rate for this process tends to be higher and shifts up substantially when the color intensity switches to gray—again, a feature unique to this student that we will discuss in Section 6.

Figure 7 summarizes the strength of the periodic component $b_j^s(t)$ as reflected by the posterior distributions of the amplitudes of each frequency. The posterior mean amplitudes are highest at frequencies 1/4 and 1/8 for all students, and most students have posterior mean amplitudes at frequency 1/8 that are higher than those at frequency 1/4. For students with large 1/8 frequency amplitudes, such as Students 3, 6 and 21, we see evidence of higher sensitivity to the stimulus distance. This signal is particularly strong for Student 6. These results suggest that it is
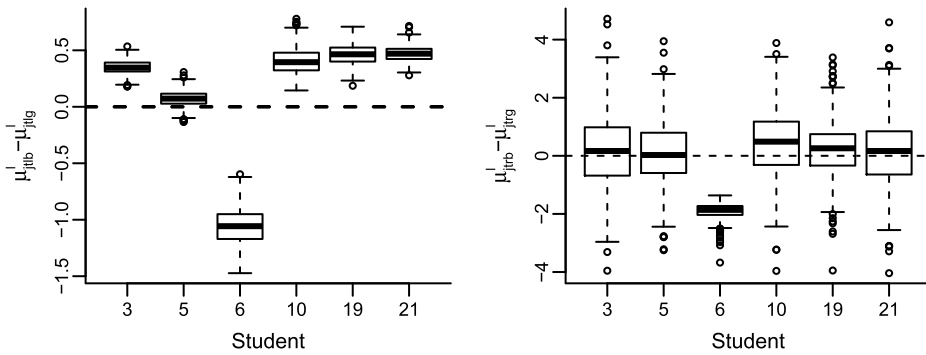
FIG. 8. *Estimates of the student-specific effect of color intensity for the correct (left) and incorrect (right) processes. The scales of the y-axes are different for the two plots.*

indeed possible to recover the "footprint" of the changing stimulus distances in the RT series.

The effect of changing stimulus color intensity in the log rates is captured by the differences $\mu^l_{jlb} - \mu^l_{jlg}$ (which equals $\mu^r_{jrb} - \mu^r_{jrg}$) for the correct process and $\mu^l_{jrb} - \mu^l_{jrg}$ (which equals $\mu^r_{jlb} - \mu^r_{jlg}$) for the incorrect process. Figure 8 shows boxplots of the posterior distribution of the color intensity effect for both processes. For the correct process all students, except Student 6, exhibit some increase in rate associated with higher color intensity, and for several students (3, 10, 19, 21) this effect is rather large. For the incorrect process, however, the effects of color intensity are negligible for all but Student 6. Consistent with what we observed earlier, Student 6 is unique in that an increase in rate is associated with *lower* color intensity.

Trend in thresholds. The estimated thresholds (posterior means and pointwise 95% credible intervals) for process s are shown in Figure 9 for selected students. The panels in the left column show the thresholds for the correct process those in the right column show the thresholds for the incorrect process. The changes in the thresholds capture several prominent local trends in the RTs shown in Figure 2.

For example, the posterior mean threshold for Student 10 is unusually high for the first 50 trials, corresponding to the student's slow responses in the same period. This elevated threshold may indicate that the student required more practice before becoming comfortable with the task. Student 6's posterior mean threshold decreases more gradually over the course of the experiment. The posterior intervals for the thresholds are wider for the incorrect than the correct process but the posterior means follow similar patterns.

Mixture parameters. Table 2 provides the posterior means of η_{Gj} and η_{Hj}, the probabilities that Student j's RTs were generated from G or H respectively as well as the posterior means of the parameters of the log-normal distribution H (a_j and b_j).
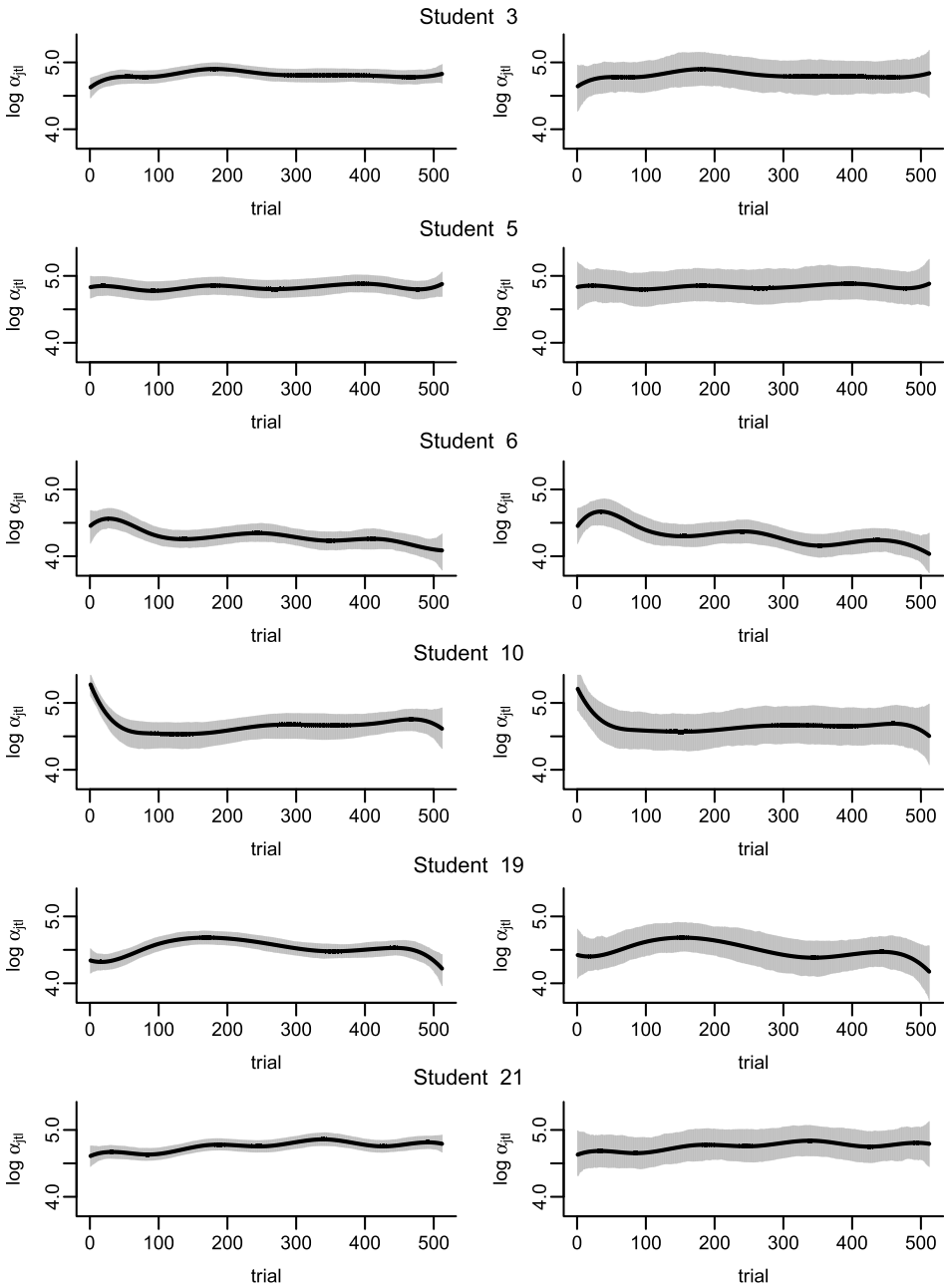
FIG. 9. *Pointwise posterior* 90% *intervals of log thresholds for selected students. The dark line is the posterior mean log threshold. The left column shows the log threshold for the correct process, and the right column shows the log thresholds for the incorrect processes.*

TABLE 2
Posterior means and (standard deviations) of the mixture parameters for selected students

	η_{Gj}	η_{Hj}	a_j	b_j
Student 3	0.030 (0.005)	0.116 (0.013)	6.140 (0.049)	0.391 (0.035)
Student 5	0.025 (0.005)	0.136 (0.014)	6.658 (0.062)	0.508 (0.038)
Student 6	0.054 (0.008)	0.180 (0.017)	6.831 (0.064)	0.499 (0.042)
Student 10	0.035 (0.006)	0.155 (0.019)	6.108 (0.084)	0.572 (0.065)
Student 19	0.029 (0.006)	0.142 (0.020)	6.675 (0.085)	0.544 (0.043)
Student 21	0.034 (0.006)	0.096 (0.012)	6.118 (0.066)	0.420 (0.047)

The estimated proportions of RTs generated from the subcognitive component G are between 0.03 and 0.05, while those from the supracognitive component H are considerably higher at 0.10–0.18. Student 6 has particularly high values of both η_{Gj} and η_{Hj}, indicating that the cognitive component of the fitted model was active less often. This reflects, to some degree, a higher proportion of trials for this student that are poorly described by the cognitive model and may be indicative of some difficulty in performing the task due to inattention or some form of physical or environmental interference. However, such interpretations should be made cautiously. We placed very low probability on proportions greater than 0.15 for both η_{Gj} and η_{Hj}, and the posterior estimates of these parameters are as much a consequence of this modeling decision as they are evidence from the data themselves. As discussed in Section 5, the results change if this feature of the prior is relaxed.

5. Model validation. We discussed our theoretical motivation for the min-inverse Gaussian distribution for RTs in Section 3. Good fits lend credibility to this model of the cognitive process. We evaluated our model fit by first looking at the coverage of predictive intervals for each stimulus condition in the students' data. Then, we compared the features of each student observed and predicted RT sequences.

We began these evaluations by drawing samples from the posterior predictive distributions $p(y_{jt}^*|\boldsymbol{y})$ for a new observation y_{jt}^* for student j at trial t. This distribution satisfies

$$p(y_{jt}^*|\boldsymbol{y}) = \int_\theta f(y_{jt}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ refers to all model parameters, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{21})$ is the full observed data for all students, $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution of $\boldsymbol{\theta}$, and $f(y_{jt}^*|\boldsymbol{\theta})$ is the mixture likelihood. We obtained Monte Carlo samples from this distribution by drawing $\boldsymbol{\theta}^b$, $b = 1, \ldots, B$, from the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{y}$ and then sampling y_{jt}^{*b} values from the mixture distribution (3.7) with parameters $\boldsymbol{\theta}^b$. Because our model assigns a different value of $\boldsymbol{\theta}$ to each trial, these predictive distributions are unique for each t.

Empirical coverage probability of 90% *predictive intervals for each stimulus condition for selected students. The number in each column header is the distance from the center (in pixels) of the stimulus and the letter g or b refers to the gray or black stimulus color intensity. The far right column gives the mean coverage across all conditions*

	$3-g$	$10-g$	$30-g$	$80-g$	$3-b$	$10-b$	$30-b$	$80-b$	**Mean**
Student 3	0.91	0.86	0.92	0.94	0.81	0.91	0.97	0.86	0.90
Student 5	0.86	0.86	0.98	0.92	0.83	0.91	0.94	0.98	0.91
Student 6	0.80	0.81	0.81	0.78	0.95	0.92	0.91	0.98	0.87
Student 10	0.94	0.92	0.83	0.73	0.86	0.92	0.95	0.94	0.89
Student 19	0.88	0.89	0.88	0.94	0.83	0.92	0.97	0.94	0.90
Student 21	0.75	0.91	0.98	0.92	0.94	0.89	0.91	0.89	0.90

Coverage of predictive intervals. We used samples from the predictive distribution to approximate predictive intervals for each student's RTs at every distance/intensity combination. We were particularly interested in verifying that the model accurately predicts changes in the RT distributions arising from changes in the stimuli. We calculated the 90% predictive intervals for each stimulus condition and then the proportion of observations that fell within those intervals. Table 3 shows the (in sample) results.

The empirical coverage is acceptable for most students for trials with the black stimulus, although it is often considerably higher than the nominal level at larger distances ($30-b$ and $80-b$). For several cells in the table the coverage is very low: Student 21 has low coverage for low-color intensity near the screen center ($3-g$), and Students 6 and 10 have low coverage for low-color intensity at a moderate distance ($30-g$). For Student 6 coverage is far below the nominal level for the conditions with the low-color intensity but is higher for the high-color intensity conditions. Some unusual features of Student 6 contribute to this phenomenon, and we will discuss a plausible explanation for it in Section 6. The other students have coverage close to the nominal level. The lower coverage proportions under these more difficult gray conditions may suggest that the model predicts RTs that are too fast for these trials. Conversely, the somewhat higher coverage proportions under the easier black conditions (especially for intermediate distances of 10 and 30 pixels) suggest that the posterior intervals may be too wide and, hence, that the model predicts too many supracognitive responses. These proportions may indicate residual misfit of the tails of the RT distributions.

Features of predicted RT series. To examine the posterior predictive RT series we generated, for each student j, 300 samples $\boldsymbol{y}_j^b = (y_{j1}^b, \ldots, y_{j512}^b), b = 1, \ldots, 300$, from that student's posterior predictive distribution. For each replication we used the same sequence of stimulus conditions experienced by the student in the original experiment. Figure 10 shows the Q–Q plots for selected students, in which we plotted the quantiles of the posterior predictive samples $\boldsymbol{y}_j^b$ (y-axis)
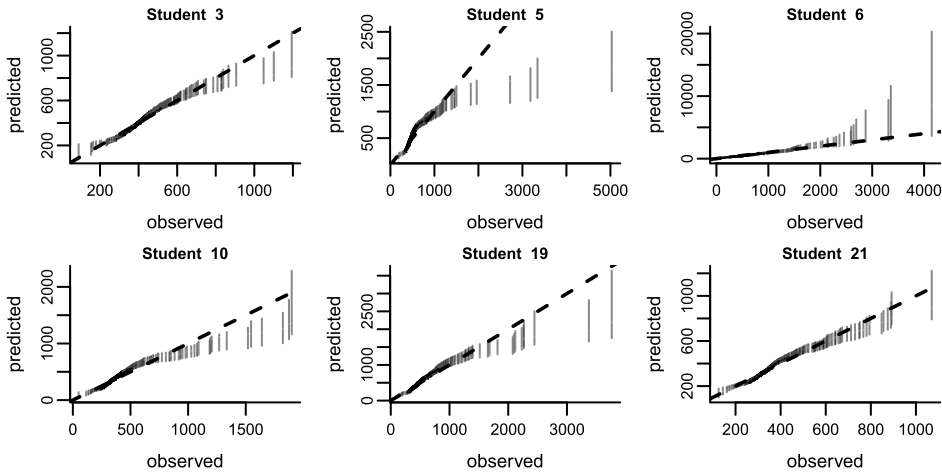
FIG. 10. *Q–Q plots of the quantiles of the posterior predictive samples vs. the quantiles of the observed RT sequences. The y = x line is the dark dashed line.*

against the quantiles of the observed sequences $\boldsymbol{y}_j$ (x-axis). Each vertical line segment represents a 95% pointwise predictive interval for that quantile. A straight line relationship between these quantiles indicates that the sequences could have realistically been drawn from the predictive distributions. To improve readability, we use different scales for the x- and y-axes.

The plots show an adequate fit for most students in most parts of the distribution. Within the central range of the RTs, the distributions of the posterior predictive RTs are similar to those that were observed, suggesting an overall good fit of the cognitive component of the model. The predictive distribution is still imperfect in the tails for some students. For Student 5, for example, the predicted quantiles are too low in both the lower and upper tails which is perhaps due to the particularly large magnitudes of this student's outliers. Students 3 and 21 show a bump at the left tail, indicating a predictive distribution that puts somewhat too much mass in the subcognitive component G. The apparent "elbows" from the preliminary fit (see Figure 4), however, are absent, a result that confirms that the inclusion of the supracognitive process H in the mixture is necessary to describe slower RTs.

In sum, the coverage proportions and the Q–Q plots show that, although the fits of the model are not perfect for all students, the mixture representation provides strong predictive performance, even in the heavy tails, without overfitting.

Predictive checks on held-out data. Another question of interest is the out-of-sample predictive performance of the model. Predicting RTs for new students not included in the data set used to fit the model is problematic because the model accounts explicitly for subject effects that vary from student to student and these effects would not be estimable for new students. It is possible, however, to fit the model on a subset of the data obtained by omitting some of the RTs for each
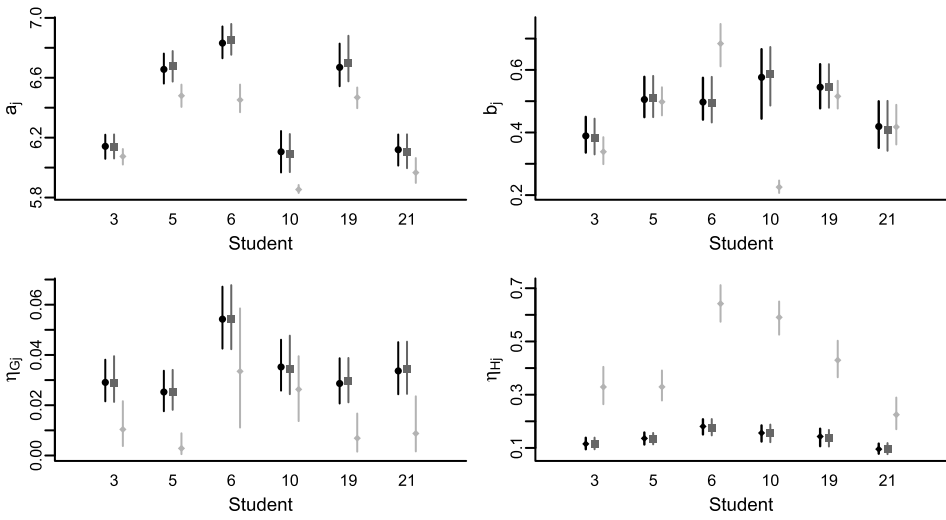
FIG. 11.   *Posterior intervals of a_j, b_j, η_{Gj} and η_{Hj} for selected students under the original priors (black lines), Prior 2 (dark gray lines) and Prior 3 (light gray lines).*

student. In this case, estimation of the student-specific effects is possible and predictions can be calculated. In the Supplementary Material we describe the results of a predictive study in which we fit the model to a subset of the original data, holding out about 17% of the trials and obtained posterior predictions of the RTs for these held-out trials. With only a few exceptions the coverage of the derived posterior predictive intervals was close to the nominal value.

Sensitivity to priors on mixture component. As we discussed in Section 3.3, we constructed the priors P_η, P_a and P_b using diagnostic information from the data; we will call these choices "Prior 1." We now evaluate the sensitivity of our results to these prior choices by fitting the model with less-informative priors that ignore the diagnostic information from the data. In particular we fit the model two additional times, changing P_a and P_b in the first run ("Prior 2") and modifying P_η in the second run ("Prior 3").

For Prior 2 we chose P_a to be log-normal$(\log(6.34), 0.15)$ and P_b to be log-normal$(\log(0.45), 0.6)$. While the prior medians of $\log(a_j)$ and $\log(b_j)$ are unchanged from Prior 1, these distributions are much more diffuse. Similarly, for Prior 3 we replaced the concentrated prior on $\boldsymbol{\eta}$ with one that has the same prior means but is much more diffuse: we used $\boldsymbol{m} = (30, 1.5, 2)$.

Figure 11 shows the posterior distributions of η_{Gj}, η_{Hj}, a_j and b_j for the original Prior 1 (black lines), Prior 2 (dark gray lines) and Prior 3 (light gray lines). None of the posterior estimates change substantially under Prior 2.

Under Prior 3 the posterior mixture probabilities change substantially. The posterior estimates of η_{Hj} are much larger, with means approximately in the range 20–65%, while the estimates of η_{Gj} are very small for most students. In addition

the values of a_j and b_j tend to decrease for most students. When η_{Hj} is large, many slower RTs that would ordinarily be attributed to the cognitive component of the model are instead attributed to the process H. These slower RTs, which are faster than the even slower "outlier" RTs, influence the log-normal parameters of H, reducing the mean of the supra-cognitive RTs and increasing the overlap between the components F and H. The more concentrated prior on η in Prior 1 prevents this substantial overlap and ensures that the mixture components are identifiable. The use of the concentrated prior is a subjective modeling choice, but one that forces the three mixture components to each play the role for which they were originally introduced.

6. Conclusions and discussion.   In this paper we have proposed a model that explains choice performance in a two-choice discrimination task. This model states that a choice between response alternatives arises as a race between competing Gaussian diffusions which implies that the pairs of RT/response data follow a min-inverse Gaussian distribution. We embedded this theoretical structure in a hierarchical framework similar to that proposed by Craigmile, Peruggia and Van Zandt (2010) in which data are modeled as a mixture of observations from a process of interest contaminated by too fast (subcognitive) and too slow (supracognitive) observations. In addition to explaining how choices (responses and RTs) are influenced by stimulus conditions, this model explains the tail behavior of the RT distributions, long-term fluctuations in the mean of the RT series and short-term dependencies across experimental trials—characteristics of RT series that are difficult for standard cognitive models to explain and that hamper accurate estimation of the effects of interest.

The min-inverse Gaussian model presented in this paper describes cognitive decisions as arising from two distinct components of the cognitive process, the diffusion rates and the decision thresholds. We modeled effects of changes in the stimuli as shifts in the log diffusion rates, while we modeled local dependencies, learning and fatigue as trends in the log decision thresholds. We collected data in an experiment in which the stimulus varied cyclically over the experimental trials producing cyclic changes in stimulus discriminability. Our goal was to isolate these cyclic changes in the log diffusion rate and to verify that the modeling framework of Craigmile, Peruggia and Van Zandt (2010) can be applied to a novel task.

The stimuli that we used varied in distance from the screen center and contrast with the screen background, stimulus features that varied systematically and repeated every eight trials. Our results showed strong effects of color intensity and distance on the diffusion rates for the process responsible for eliciting a correct response; these effects were consistent with increasing diffusion rates under easy discrimination conditions and decreasing rates under difficult conditions. Effects of stimulus color intensity were stronger and more regular than those of stimulus distance, but the posterior estimates of the amplitudes of different frequencies

in a harmonic regression on the log diffusion rates were consistent with the pattern of changes in stimulus discriminability. No strong effects of color intensity were detected for the process controlling the incorrect response, possibly due to the paucity of incorrect responses for most students.

Student 6 is the only exception to the preceding comments. The most striking features uncovered by the analysis are that, for the black color intensity, the response rates of Student 6 tend to be lower than those of the other students and that the effect of color intensity is reversed. An *ex-post-facto* exploratory analysis revealed that the error rate for this student is much higher than for the other students. In particular Student 6's performance when the stimulus is gray is no better than guessing. This suggests that, possibly due to an unreported vision deficiency, Student 6 did not perform the required task in the low-contrast condition but rather resorted to rapid guessing. The student, however, appeared to be able to perform the task adequately under the high-contrast condition, albeit more slowly than the other students. This ability to uncover and describe unusual behavior in the experiment participants is a testament to the flexibility and discriminating capabilities of the model and its potential for fruitful applications in diagnostic settings in which discovery and characterization of idiosyncratic features is necessary.

There are many other alternative modeling approaches that could describe RT behavior, accommodating the changes in response rates due to experimental design, fatigue, and learning. In the Supplementary Material we consider one such alternative based on a generalized additive model (GAM) and compare its estimates and predictive performance to those of our proposed model. In many respects the two approaches perform similarly, but, in our view, our approach has two fundamental advantages. First, the min-inverse Gaussian model is grounded in an established theory of cognitive processing, and, as a by-product, its estimates can be meaningfully interpreted. Second, our approach models jointly the response values and the RTs while the GAM only describes the RTs marginally, with no obvious path toward joint modeling.

One novelty introduced in this paper is the use of diagnostic information from the data to inform the priors on the mixture model parameters. Inclusion of the sub- and supracognitive mixture components improved the model fit substantially compared to a preliminary fit using only the cognitive model. This finding is consistent with our earlier work arguing that these mixture components are necessary for RT data [Craigmile, Peruggia and Van Zandt (2010), Kim et al. (2017)]. The hierarchical nature of our model also allows for substructures in the model that can predict extremely fast and slow RTs adequately for different individuals. The mixture representation mitigates the influence of contaminant observations so that the effects of stimulus changes may be estimated more accurately.

While our model is tailored to the design of our experiment, the proposed framework is easily generalized to future experiments by small modifications to the hierarchical models on the log rates or log thresholds. For example, additional factors may be accommodated in the baseline log rate, and the cyclic component may be

omitted in experiments without a repeating signal. Our model for the log thresholds does not depend on experimental design and may readily be applied to any experimental design or modified to allow more or less local flexibility. Different choices of prior distributions on the parameters of the mixture component may be appropriate for new datasets, and these can be determined using the same principles outlined in Section 3.3. Most importantly, when applied to a future populations, this framework has the potential to isolate experimental effects across individuals that share common characteristics and allow for individual differences within groups.

SUPPLEMENTARY MATERIAL

Supplement A (DOI: 10.1214/18-AOAS1192SUPP; .pdf). This supplement provides additional detail on the experimental procedure, displays the RT sequences for all students, and reports the results of the analysis for all students. This supplement also includes a demonstration of the ability of our estimation procedure to recover the parameter values used to generate simulated data for a prototypical experimental participant, a comparison of the performance of our *theoretically-motivated* modeling framework with that of a *descriptive* generalized additive model, and an evaluation of the predictive performance of our approach.

REFERENCES

BAAYEN, R. H. and MILIN, P. (2010). Analyzing reaction times. *International Journal of Psychological Research* **3** 12–28.

BROWN, S. D. and HEATHCOTE, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation **57** 153–178.

CAPLIN, A. and MARTIN, D. (2016). The dual-process drift diffusion model: Evidence from response times. *Economic Inquiry* **54** 1274–1282.

CRAIGMILE, P. F., PERUGGIA, M. and VAN ZANDT, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika* **75** 613–632. MR2741490

HEITZ, R. P. and SCHALL, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron* **76** 616–628.

KIM, S., POTTER, K., CRAIGMILE, P. F., PERUGGIA, M. and VAN ZANDT, T. (2017). A Bayesian race model for recognition memory. *J. Amer. Statist. Assoc.* **112** 77–91. MR3646554

KUNKEL, D., POTTER, K., CRAIGMILE, P. F., PERUGGIA, M. and VAN ZANDT, T. (2019). Supplement to "A Bayesian race model for response times under cyclic stimulus discriminability." DOI:10.1214/18-AOAS1192SUPP.

LOGAN, G. D., VAN ZANDT, T., VERBRUGGEN, F. and WAGENMAKERS, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review* **121** 66–95.

LUCE, R. D. (1986). *Response Times*: *Their Role in Inferring Elementary Mental Organization*. Oxford Univ. Press, Oxford, UK.

NELSON, M. J., MURTHY, A. and SCHALL, J. D. (2016). Neural control of visual search by frontal eye field: Chronometry of neural events and race model processes. *J. Neurophysiol.* **115** 1954–1969.

RATCLIFF, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* **114** 510–532.

RATCLIFF, R. and MCKOON, G. (2007). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* **20** 873–922.

RATCLIFF, R., SMITH, P. L. and MCKOON, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Curr. Dir. Psychol. Sci.* **24** 458–470.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

USHER, M. and MCCLELLAND, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review* **108** 550–592.

VANDEKERCKHOVE, J. and TUERLINCKX, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychon. Bull. Rev.* **14** 1011–1026.

VAN ZANDT, T., COLONIUS, H. and PROCTOR, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychon. Bull. Rev.* **7** 208–256.

WAGENMAKERS, E.-J., FARRELL, S. and RATCLIFF, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychon. Bull. Rev.* **11** 579–615.

WHELAN, R. (2008). Effective analysis of reaction time data. *Psychological Record* **58** 475–482.

D. KUNKEL
P. F. CRAIGMILE
M. PERUGGIA
DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO
USA
E-MAIL: kunkel.48@osu.edu
        pfc@stat.osu.edu
        peruggia@stat.osu.edu

K. POTTER
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MASSACHUSETTS
AMHERST, MASSACHUSETTS
USA
E-MAIL: kevin.w.potter@gmail.com

T. VAN ZANDT
DEPARTMENT OF PSYCHOLOGY
OHIO STATE UNIVERSITY
COLUMBUS, OHIO
USA
E-MAIL: van-zandt.2@osu.edu