Behavior Research Methods (2020) 52:1991–2007
https://doi.org/10.3758/s13428-020-01368-6

Modeling pupil responses to rapid sequential events

Rachel N. Denison[1] · Jacob A. Parker[1,2] · Marisa Carrasco[1]

Published online: 6 March 2020

Abstract

Pupil size is an easily accessible, noninvasive online indicator of various perceptual and cognitive processes. Pupil measurements have the potential to reveal continuous processing dynamics throughout an experimental trial, including anticipatory responses. However, the relatively sluggish (~2 s) response dynamics of pupil dilation make it challenging to connect changes in pupil size to events occurring close together in time. Researchers have used models to link changes in pupil size to specific trial events, but such methods have not been systematically evaluated. Here we developed and evaluated a general linear model (GLM) pipeline that estimates pupillary responses to multiple rapid events within an experimental trial. We evaluated the modeling approach using a sample dataset in which multiple sequential stimuli were presented within 2-s trials. We found: (1) Model fits improved when the pupil impulse response function (PuRF) was fit for each observer. PuRFs varied substantially across individuals but were consistent for each individual. (2) Model fits also improved when pupil responses were not assumed to occur simultaneously with their associated trial events, but could have non-zero latencies. For example, pupil responses could anticipate predictable trial events. (3) Parameter recovery confirmed the validity of the fitting procedures, and we quantified the reliability of the parameter estimates for our sample dataset. (4) A cognitive task manipulation modulated pupil response amplitude. We provide our pupil analysis pipeline as open-source software (Pupil Response Estimation Toolbox: PRET) to facilitate the estimation of pupil responses and the evaluation of the estimates in other datasets.

Keywords Pupillometry · Model · Pupil response function · Temporal attention · PRET

Introduction

Pupil size depends strongly on light levels, but it also covaries with an array of perceptual and cognitive processes—from attention to memory to decision making (for recent reviews, see Binda & Gamlin, 2017; Ebitz & Moore, 2018; Mathôt, 2018; Wang & Munoz, 2015). Pupil size can be measured noninvasively and continuously, making pupillometry a promising tool for probing the ongoing dynamics linked to these processes. The pupil dilates in response to task-relevant stimuli (Hoeks & Levelt, 1993; Kang, Huffer, & Wheatley, 2014; Wierda, van Rijn, Taatgen, & Martens, 2012; Willems, Damsma, Wierda, Taatgen, & Martens, 2015; Willems, Herdzin, & Martens, 2015; Zylberberg, Oliva, & Sigman, 2012) and arousing, interesting, or surprising stimuli (Allen et al., 2016; Hess & Polt, 1960; Kloosterman et al., 2015; Knapen et al., 2016; Libby, Lacey, & Lacey, 1973; Nassar et al., 2012; Preuschoff, 't Hart, & Einhäuser, 2011), as well as in concert with internally driven cognitive events, like mental calculation (Hess & Polt, 1964; Kahneman, Beatty, & Pollack, 1967), memorization (Kahneman & Beatty, 1966; Kang et al., 2014), and decision formation (Cheadle et al., 2014; de Gee et al., 2017; de Gee, Knapen, & Donner, 2014; Lempert, Chen, & Fleming, 2015; Murphy, Boonstra, & Nieuwenhuis, 2016; Murphy, Vandekerckhove, & Nieuwenhuis, 2014; Urai, Braun, & Donner, 2017; van Kempen et al., 2019). Pupil dilation is mediated by activity in the locus coeruleus (LC), hypothalamus, and superior colliculus, which interact with the pathways that control pupillary dilation and constriction (Mathôt, 2018; Wang & Munoz, 2015). The pupil time series may therefore carry information about multiple events within an

Rachel N. Denison and Jacob A. Parker contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-020-01368-6) contains supplementary material, which is available to authorized users.

✉ Rachel N. Denison
rachel.denison@nyu.edu

[1] Department of Psychology and Center for Neural Science, New York University, 6 Washington Place, New York, NY 10003, USA

[2] Human Motor Control Section, Medical Neurology Branch, National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD, USA

experimental trial, as well as about anticipatory neural responses not available from behavioral reports alone, which provide retrospective rather than online measures.

A critical challenge in relating pupillary changes to specific perceptual and cognitive processes is that pupillary dynamics are relatively slow. Whereas perception and cognition unfold over timescales of a few hundred milliseconds, the pupil takes about 2 s to dilate and return to baseline in response to a single, brief stimulus (Hoeks & Levelt, 1993). However, the limiting factor in the speed of pupil dilation does not appear to be the dynamics of the neural responses that drive the pupil. For example, LC activity is tightly linked to pupil dilation (Aston-Jones & Cohen, 2005; de Gee et al., 2017; Joshi, Li, Kalwani, & Gold, 2016; Murphy, O'Connell, O'Sullivan, Robertson, & Balsters, 2014; Reimer et al., 2016; Varazzani, San-Galli, Gilardeau, & Bouret, 2015) and has much faster dynamics. LC neurons fire with a latency of ≤ 100 ms after a task-relevant stimulus, with a brief, phasic response (Aston-Jones & Cohen, 2005; Foote, Aston-Jones, & Bloom, 1980; Sara & Bouret, 2012). Therefore, the pupil size at a given time may reflect the influence of multiple preceding or ongoing internal signals related to distinct perceptual and cognitive processes. The standard approach to pupillometry, namely measuring the pupil size time series, cannot disentangle the influences of these various signals on the pupil size.

To address this challenge, researchers have begun to use models to link changes in pupil size to the distinct internal signals elicited by specific trial events (de Gee et al., 2017; de Gee et al., 2014; Hoeks & Levelt, 1993; Johansson & Balkenius, 2017; Kang et al., 2014; Kang & Wheatley, 2015; Knapen et al., 2016; Korn & Bach, 2016; Korn, Staib, Tzovara, Castegnetti, & Bach, 2017; Lempert et al., 2015; Murphy et al., 2016; Urai et al., 2017; van den Brink, Murphy, & Nieuwenhuis, 2016; van Kempen et al., 2019; Wierda et al., 2012; Willems, Damsma, et al., 2015; Willems, Herdzin, & Martens, 2015; Zylberberg et al., 2012). These signals can be thought of as the internal responses to trial events that drive pupil dilation, and the goal of modeling is to infer the properties of these internal signals, such as their amplitudes, from the pupil time series. Under constant luminance conditions, it is typical to model pupil dilations only, which are considered to be linked to internal signals that drive the sympathetic pupillary pathway (Mathôt, 2018).

Pupil response models typically incorporate two principles based on the work of Hoeks and Levelt (1993). First, the models assume a stereotyped pupil response function (PuRF), which describes the time series of pupil dilation in response to a brief event. These authors found that the PuRF is well described by a gamma function, and they reported average parameters for that function, which have been used in many studies (de Gee et al., 2017; de Gee et al., 2014; Kang et al., 2014; Kang & Wheatley, 2015; Lempert et al., 2015; Murphy et al., 2016; van Kempen et al., 2019; Wierda et al., 2012; Willems, Damsma, et al., 2015; Willems, Herdzin, & Martens, 2015; Zylberberg et al., 2012)—we refer to this specific form of the PuRF as the "canonical PuRF". Second, the models assume that pupil responses to different trial events sum linearly to generate the pupil size time series; that is, they are general linear models (GLMs). This assumption is based on Hoeks and Levelt's (1993) finding that, for the tested stimulus parameters, the pupil responded like a linear system. Incorporating these two principles, the pupil response to sequential trial events has been modeled as the sum of component pupil responses, where each component response is the internal signal time series associated with a single trial event convolved with the PuRF. Using this approach, the pupil has been found to track decision periods (de Gee et al., 2017; de Gee et al., 2014; Lempert et al., 2015; Murphy et al., 2016; Murphy, Vandekerckhove, & Nieuwenhuis, 2014; van Kempen et al., 2019) and fluctuations in target identification during a rapid stimulus sequence (Wierda et al., 2012; Zylberberg et al., 2012)—findings that reveal the faster internal dynamics underlying the measured pupil time series.

Despite the promise of using such models to link distinct pupillary responses to specific trial events, there is currently no standard procedure for modeling the pupil time series. Hoeks and Levelt (1993) estimated the number, timing, and amplitudes of impulse signals that drove pupil dilations. Later studies assumed that every stimulus presentation was associated with a concurrent impulse signal and estimated only their amplitudes. Some studies have also included longer, cognitive events, like a decision period (de Gee et al., 2017; de Gee et al., 2014; Lempert et al., 2015; Murphy et al., 2016; van Kempen et al., 2019), or extra parameters to account for slow drifts in pupil size across the trial (Kang et al., 2014; Kang & Wheatley, 2015; Wierda et al., 2012; Willems, Damsma, et al., 2015; Willems, Herdzin, & Martens, 2015; Zylberberg et al., 2012). Most studies have used the canonical PuRF (de Gee et al., 2017; de Gee et al., 2014; Kang et al., 2014; Kang & Wheatley, 2015; Lempert et al., 2015; Murphy et al., 2016; van Kempen et al., 2019; Wierda et al., 2012; Willems, Damsma, et al., 2015; Willems, Herdzin, & Martens, 2015; Zylberberg et al., 2012), but others used more complicated PuRFs that could also capture pupil dilations and constrictions in response to changes in illumination (Korn & Bach, 2016; Korn et al., 2017), or that separately modeled transient and sustained components of the pupil dilation (Spitschan et al., 2017). Importantly, these pupil-modeling methods have not been systematically evaluated or compared, hindering the adoption of a field-wide standard.

Here we evaluate GLM procedures for modeling the pupil time series for trials with multiple rapid sequential events, under constant illumination. We conduct factorial model comparison to determine which parameters should be included, and we perform several validation and reliability tests using the best model. Based on the results, we recommend a specific

model structure and fitting procedure for more general adoption and future testing. Critically, we found that timing parameters not typically included in pupil GLMs substantially improve model fits. Our findings indicate that PuRF timing should be estimated for each observer, rather than assuming the canonical PuRF, and that the internal signals driving the pupil response are not necessarily concurrent with stimulus onsets. We provide an open-source MATLAB toolbox, the Pupil Response Estimation Toolbox (PRET), which fits pupil GLMs to obtain event-related amplitudes and latencies, estimates parameter reliabilities, and compares models.

As a case study, here we analyzed data for a study on temporal attention – the prioritization of sensory information at specific points in time (Denison, Heeger, & Carrasco, 2017). Combining information about the expected timing of sensory events with ongoing task goals improves our perception and behavior (review by Nobre & van Ede, 2018). By studying the effects of temporal attention on perception, we can better understand the dynamics of visual perception. To understand these dynamics, a critical distinction must be made between temporal attention—prioritization of task-relevant time points—and temporal expectation—prediction of stimulus timing regardless of task relevance. Here, we manipulated temporal attention while equating expectation by using precues to direct voluntary temporal attention to specific stimuli in predictably timed sequences of brief visual targets (Denison et al., 2017; Denison, Yuval-Greenberg, & Carrasco, 2019; Fernandez, Denison, & Carrasco, 2019). Within the temporal attention dataset, we also compared two kinds of tasks—orientation discrimination and orientation estimation—which involved identical stimulus sequences and only differed in the required report. It is likely that estimation has a higher cognitive demand than discrimination, as it requires a precise response, as opposed to a two-alternative forced choice. Thus, the physical stimuli were fixed while the cognitive demand varied between tasks. This dataset provided a good case study to evaluate GLM procedures for modeling the pupil time series as it had multiple rapid sequential events, required temporally precise cognitive control to attend to a relevant time point that varied from trial to trial, and included an orthogonal task manipulation that involved different cognitive demands for identical stimuli.

Methods

Data set

We reanalyzed eye-tracking data collected in a recent study on temporal attention by Denison et al. (2017). Thus, behavioral procedures were identical to those previously reported (Denison et al., 2017; Denison et al., 2019). To maximize the power of the pupil analysis, we combined the data from the two experiments in that study with identical stimulus sequences (Experiments 1 and 3). Experiment 1 used an orientation discrimination task, so we refer to it here as the Discrimination experiment. Experiment 3 used an orientation estimation task, so we refer to it here as the Estimation experiment. The stimuli were similar across experiments: on each trial, human observers were presented with a predictably timed sequence of two target gratings—which we refer to as T1 and T2—and judged the orientation of one of these gratings. An auditory precue before each sequence directed temporal attention to one or both grating times, and an auditory response cue after each sequence instructed observers which grating's orientation to report (see Fig. 1 and Behavioral procedures).

Observers

The observers were the same as in Denison et al. (2017), except that eye-tracking data from four observers in the Discrimination experiment could not be used for pupil analysis for technical reasons (e.g., excessive blinking: blink overlapped with response cue in > 20% of trials). To better equate the number of observers in each experiment for the present study, we collected data from three new observers for the Discrimination experiment (as also reported in Denison et al., 2019). This gave 21 total observer data sets: nine in Discrimination and 12 in Estimation. Three observers participated in both experiments, including author R.N.D. Therefore, 18 unique observers (ten female, eight male; aged 19–43 years) are included in the present study. All observers provided informed consent, and the University Committee on Activities involving Human Subjects at New York University approved the experimental protocols. All observers had normal or corrected-to-normal vision.

Stimuli

Stimuli were generated on an Apple iMac using MATLAB (The MathWorks, Inc., Natick, MA, United States) and Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). They were displayed on a gamma-corrected Sony Trinitron G520 CRT monitor with a refresh rate of 100 Hz at a viewing distance of 56 cm. Observers' heads were stabilized by a chin-and-head rest. A central white fixation "x" subtended 0.5° visual angle. Visual target stimuli were 4 cpd sinusoidal gratings with a 2D Gaussian spatial envelope (standard deviation 0.7°), presented in the lower right quadrant of the display centered at 5.7° eccentricity (Fig. 1a). Stimuli were high contrast (64% or 100%, which we combined as there were no behavioral differences). Placeholders, corners of a 4.25° x 4.25° white square outline (line width 0.08°) centered on the target location, were present throughout the display to minimize spatial uncertainty. The stimuli were presented on a medium gray background (57 cd/m2). Auditory precues were high (precue

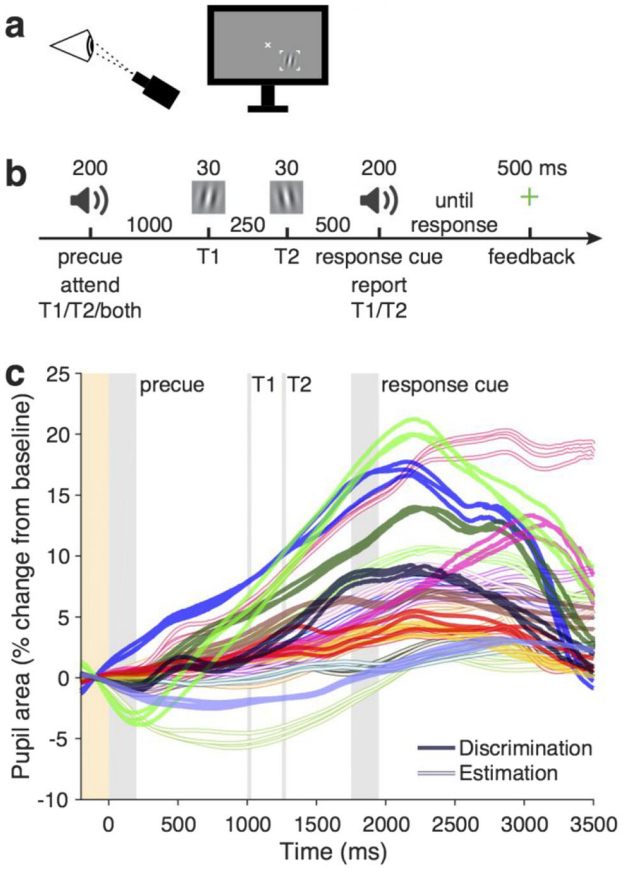

Fig. 1 Task and pupil time series. **a** Setup of visual display and eye tracker. **b** Trial sequence. In the Discrimination task, observers reported whether the target stimulus was tilted CW or CCW. In the Estimation task, a grating probe (not shown) appeared after the response cue, and observers adjusted it to report the exact orientation of the target stimulus. **c** Pupil time series (*colored lines*), mean across trials in each condition for each observer. *Filled lines* are observers in the Discrimination experiment, and *open lines* are observers in the Estimation experiment. Each observer has a unique color, and three observers participated in both experiments (same color filled and empty). Three lines per observer and experiment show different precuing conditions (precue T1, T2, neutral), i.e., independent sets of trials. Time series were baseline-normalized per trial (*baseline period shaded yellow*). *Gray shaded regions* are trial events.

T1: 784 Hz; G5) or low (precue T2: 523 Hz; C5) pure sine wave tones, or their combination (neutral precue). Auditory stimuli were presented on the computer speakers.

Behavioral procedures

Basic task and trial sequence Observers judged the orientation of grating patches that appeared in short sequences of two target stimuli per trial (T1 and T2). Targets were presented for 30 ms each at the same spatial location, separated by stimulus onset asynchronies (SOAs) of 250 ms (Fig. 1b). An auditory precue 1000 ms before the first target instructed observers to attend to one or both of the targets. Thus, there were three precue types: attend to T1, attend to T2, or attend to both targets. Observers were asked to report the orientation of one of the targets, which was indicated by an auditory response cue 500 ms after the last target. The duration of the precue and response cue tones was 200 ms. The timing of auditory and visual events was the same on every trial.

Discrimination task In the Discrimination experiment, observers performed an orientation discrimination task (Fig. 1b). Each target was tilted slightly clockwise (CW) or counterclockwise (CCW) from either the vertical or horizontal axis, with independent tilts and axes for each target. Observers pressed a key to report the tilt (CW or CCW) of the target indicated by the response cue, with unlimited time to respond. Tilt magnitudes were determined separately for each observer by a thresholding procedure before the main experiment. Observers received feedback at fixation (correct: green "+"; incorrect: red "-") after each trial, as well as feedback about performance accuracy (percent correct) following each experimental block.

Estimation task In the Estimation experiment, observers performed an orientation estimation task (Fig. 1b). Target orientations were selected randomly and uniformly from 0 to 180°, with independent orientations for each target. Observers estimated the orientation of the target indicated by the response cue by adjusting a grating probe to match the perceived target orientation. The probe was identical to the target but appeared in a new random orientation. Observers moved the mouse horizontally to adjust the orientation of the probe and clicked the mouse to submit the response, with unlimited time to respond. The absolute difference between the reported and presented target orientation was the error for that trial. Observers received feedback at fixation after each trial (error < 5°, green "+"; 5–10°, yellow "+"; ≥ 10°, red "-"). Additional feedback after each block showed the percent of trials with < 5° errors, which were defined to observers as "correct".

Training and testing sessions All observers completed one session of training prior to the experiment to familiarize them with the task and, in the Discrimination experiment, determine their tilt thresholds. Thresholds were selected to achieve ~79% performance on neutral trials. Observers completed 640 trials across two 1-h sessions. All experimental conditions were randomly interleaved across trials.

Eye data collection

Pupil size was continuously recorded during the task at a sampling frequency of 1000 Hz using an EyeLink 1000 eye tracker (SR Research). Raw gaze positions were converted into degrees of visual angle using the 5-point-grid calibration, which was performed at the start of each experimental run. Online streaming of gaze positions was used to ensure central

fixation (< 1.5° from the fixation cross center) throughout the experiment. Initiation of each trial was contingent on fixation, with a 750-ms minimum inter-trial interval. Observers were required to maintain fixation, without blinking, from the onset of the precue until 120 ms before the onset of the response cue. If observers broke fixation during this period, the trial was stopped and repeated at the end of the block.

Preprocessing

Data files from the eye tracker were imported to MATLAB to perform all preprocessing and modeling with custom software (Pupil Response Estimation Toolbox, PRET). The raw time series from each session was epoched into trials spanning from –500 to 3500 ms, relative to the precue at 0 ms. Blinks were interpolated trial by trial using a cubic spline interpolation method (Mathôt, 2013). All trials were individually baseline normalized by calculating the average pupil size over the region from –200 to 0 ms, then calculating the percent difference from this baseline at each point along the time series:

$$x_{norm} = \frac{x - \text{baseline}}{\text{baseline}} \times 100\%, \quad (1)$$

where x_{norm} is the normalized data and x is the raw data. We normalized the time series in this way to obtain meaningful units of percent change from baseline, but we note there are also arguments for a purely subtractive baseline correction procedure (Mathôt, Fabius, Heusden, & Stigchel, 2018; Reilly, Kelly, Kim, Jett, & Zuckerman, 2018). Trials were grouped into conditions depending on the precue (T1, T2, neutral), and the mean time series was calculated across trials in each condition for each observer.

Pupil modeling

GLM modeling framework Measured pupil size time series were modeled as a linear combination of component pupil responses (Fig. 2). A component pupil response is the predicted pupil size time series associated with a single internal (neural) signal that leads to pupil dilation. Mathematically, an internal signal was represented as a time series concurrent with the measured pupil size time series. The component pupil response for a given internal signal was calculated by convolving the signal time series with a PuRF. The general PuRF takes the form

$$h(t) = t^n e^{-nt/t_{\max}}, \quad (2)$$

where h is the pupil size, t is the time in ms, n controls the shape of the function, and $t_{\max}$ controls the temporal scale of the function and is the time of the maximum (Hoeks & Levelt, 1993) (Fig. 2a). For a given measured pupil size time series, each internal signal was convolved with the same PuRF. Each component pupil response was assumed to be dilatory.

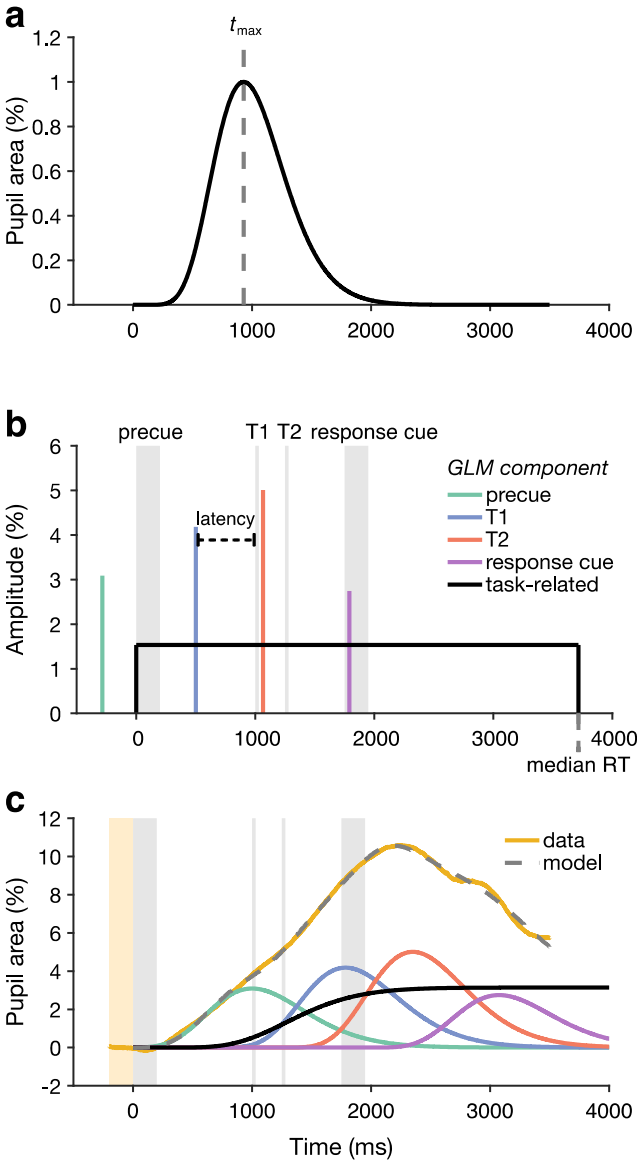

Fig. 2 General linear modeling of the pupil time series. **a** Pupil response function (PuRF), which describes the pupillary response to an impulse event. The canonical PuRF, an Erlang gamma function with $n = 10.1$ and $t_{max} = 930$ ms (*vertical dashed line*), is shown. **b** Internal signals hypothesized to drive pupil dilation. The internal signal associated with each trial event (brief auditory and visual stimuli, *gray shaded regions*) is modeled as a delta function (*vertical colored lines*) with some amplitude and some latency with respect to the event. A sustained, task-related signal could also be modeled. Shown here is a boxcar (*black line*), which starts at the onset of the precue and lasts until the median RT of the modeled trials. **c** The mean pupil time series across trials (*yellow line*) is modeled in two steps. First, each internal signal time series is convolved with the PuRF to form component pupil responses (*colored lines*, legend in panel **b**). Second, the component pupil responses are summed to calculate the model prediction (*gray dashed line*). Parameters of the model, such as the amplitudes and latencies of the internal event signals, are fit using an optimization procedure

We assumed there was a transient internal signal (de Gee et al., 2014; Hoeks & Levelt, 1993; Wierda et al., 2012) associated with each event in the trial sequence: the precue, T1, T2, and the response cue. Each of these event-related signals took the form of a Dirac delta function. An additional internal signal could be included to model a constant, sustained signal associated with task engagement (de Gee et al., 2014). This task-related signal took the form of a boxcar function, with nonzero values starting at the precue and ending at the median response time of the observer being modeled. Thus, we modeled our measured pupil size time series as the linear combination of up to four event-related and one task-related component pupil responses.

Model parameters We fit models of up to 11 parameters to a given pupil size time series. The possible parameters were: internal signal amplitudes and latencies for each trial event; internal signal amplitude for the task-related response or alternatively a slope parameter specifying a linear drift across the trial; one parameter specifying the timing of the PuRF; and one baseline shift parameter. The details were as follows:

Each event-related signal had an amplitude parameter and a latency parameter associated with it. The amplitude parameter was the value of the nonzero point of the delta function and indicated the magnitude of the internal signal, and thus determined the magnitude of the component pupil response associated with it. The latency parameter was the time (in ms) of the nonzero value, relative to the time of its corresponding event. The pupil latency could be positive, after the event, or negative, before the event. As the timing of the stimuli was perfectly predictable—observers knew in advance when the stimuli would appear—we allowed for the possibility that observers could start attending before the stimulus appeared, driving pupil dilation in advance of the stimulus. The task-related signal only had an amplitude parameter associated with it because it was assumed to start at the beginning of the trial.

The PuRF that was convolved with each signal had two parameters: t_{max}, which controls the temporal scale and time of the peak, and n, which controls the shape of the function. Only t_{max} was estimated while n was set to the canonical value of 10.1 (Hoeks & Levelt, 1993). The t_{max} parameter can be interpreted as the time it takes an observer's pupil to dilate maximally in response to an internal signal. The PuRF was normalized such that the event-related and task-related amplitude parameters indicated the percentage increase in pupil size attributable to the corresponding signal. The PuRF was normalized to a maximum value of 1, so that an amplitude value of 1 corresponded to a 1% increase in pupil size from baseline. For the task-related amplitude, the PuRF was normalized such that the PuRF convolved with the boxcar had a maximum value of 1. Thus, a task-related amplitude of 1 also corresponded to a 1% increase in pupil size from baseline to peak size.

The final parameter was a baseline shift parameter we termed the y-intercept (y-int). The y-int parameter was simply a shift along the y-axis of the entire predicted pupil size time series. We included this in the model because we noticed that for some observers, although all trials were baseline-normalized during preprocessing based on a time window before the precue, pupil size was decreasing during this window and continued to decrease until shortly after the precue. This meant that pupil dilations during the trial sequence started from a value below the calculated baseline. Without accounting for this shift in baseline with the y-int parameter, the model would underestimate the amplitude of the component pupil responses.

Model comparison and selection Previous linear models based on the PuRF only estimated the amplitude of component pupil responses (e.g., de Gee et al., 2014; Wierda et al., 2012). The component responses were assumed to onset at the time of the corresponding trial event, and PuRFs were assumed to be identical across observers. However, these assumptions have never been systematically evaluated. Here, we asked whether introducing additional timing parameters would allow more accurate modeling of the pupil response. In addition, the characteristic slow pupil dilation throughout a trial has been modeled in different ways, as either a linear drift (Wierda et al., 2012) or a sustained task-related response convolved with the PuRF (de Gee et al., 2014). We compared these two possibilities.

We compared 24 different models (Table 1). These models included all permutations of latency, t_{max}, and y-int as fixed vs. free parameters and three different forms of the task-related component. Fixed values are reported in Table 1. The three task-related components tested were a boxcar function convolved with the PuRF, a linear function, and no task-related component. The boxcar function was nonzero from 0 ms (the onset of the precue) to the median response time for a given observer and had one amplitude parameter for the height of the boxcar. The linear function had a slope parameter and a fixed intercept of 0%. It represented a linear drift throughout the whole trial and did not depend on response time. Amplitude parameters for each trial event were always estimated. Each model was fit to the mean time series for each condition and observer. We also checked that the results held when fitting to the single trial time series; in this case, baseline-corrected single trial time series were concatenated and fit to concatenated model time series.

To compare models, the Bayesian information criterion (BIC) was calculated for each model and observer across conditions and averaged at the group level to get one metric per model. The model with the lowest metric for most observers

Table 1 Factorial model comparison

Model #	Model code	Amplitude	Latency	*y*-int	t_{max}	Task-related	# params	ΔBIC
1	LYT-B					Boxcar (1)	11	0
2	LYT-L					Linear (1)	11	-49
3	LYT-0				Parameter (1)	None	10	3,328
4	LY-B					Boxcar	10	7,939
5	LY-L					Linear	10	7,006
6	LY-0			Parameter (1)	930 ms	None	9	19,196
7	LT-B					Boxcar	10	8,840
8	LT-L					Linear	10	10,946
9	LT-0				Parameter	None	9	12,344
10	L-B					Boxcar	9	14,069
11	L-L					Linear	9	14,609
12	L-0		Parameter (4)	0%	930 ms	None	8	25,178
13	YT-B					Boxcar	7	14,688
14	YT-L					Linear	7	20,564
15	YT-0				Parameter	None	6	20,296
16	Y-B					Boxcar	6	27,319
17	Y-L					Linear	6	26,392
18	Y-0			Parameter	930 ms	None	5	40,876
19	T-B					Boxcar	6	22,556
20	T-L					Linear	6	25,720
21	T-0				Parameter	None	5	27,568
22	0-B					Boxcar	5	30,585
23	0-L					Linear	5	31,841
24	0-0	Parameter (4)	0 ms	0%	930 ms	None	4	45,679

All models tested are shown with their parameters. *Rows* are models; *columns including gray shading* are factors. *Gray shading* indicates parameters were fit, with number of parameters in parentheses. *No shading* indicates parameters were fixed, with the fixed value given. Model code: L = latency, Y = *y*-int, T = t_{max}, 0 = none; task component: B = boxcar, L = linear, 0 = none

was selected as the best model and used for further analysis. The BIC was chosen as the comparison metric to account for the differing numbers of parameters among models.

We also assessed cross-validated R^2 for each model using a fourfold cross-validation procedure, in which models were fit to 75% of the data and tested on the remaining 25%. Cross-validated R^2 was calculated by comparing the model prediction to the mean across trials of the held-out data for each fold and averaging across folds. A noise ceiling for R^2 was calculated by comparing the mean across trials of the fitted data to the mean across trials of the held-out data for each fold and averaging across folds. R^2 values were computed for each model for each observer and then averaged across observers.

Parameter estimation Model parameters were estimated for the mean pupil time series for each condition and observer using a two-phase procedure. In both phases, the cost function for determining goodness of fit was defined as the sum of the squared errors between the measured pupil time series and the predicted (model-generated) pupil time series in the time window from 0 to 3500 ms. First, the cost function was evaluated at 2000 sets of parameter values sampled from independent uniform distributions of each parameter. These distributions were bounded by the parameter constraints described below. Second, constrained optimizations (MATLAB *fmincon*) were performed starting from the 40 sets of parameter values with the lowest cost from the first phase. The set of optimized parameter values minimizing the cost function was selected as the best estimate.

Parameter constraints were selected to ensure parameters could vary meaningfully within physiologically feasible ranges. Event and task-related amplitudes were constrained to the range of 0 to 100, meaning that any single internal signal could not evoke a pupil response in excess of 100% change from baseline. This range also enforces that pupil responses refer to dilation as opposed to constriction. Event-related latencies were constrained to a range of –500 to 500 ms. The t_{max} value was constrained to a range of 0 to 2000 ms. The *y*-int parameter was constrained to a range of –20 to 20% change from baseline. The slope parameter during model comparison was constrained to a range of 0 to 50% change over the trial window of 3500 ms.

Bootstrapping procedure for parameter estimation To obtain robust parameter estimates, and to quantify the uncertainty in the parameter estimates due to both noise in the data and variability in the fitting procedure, we used bootstrapping. Parameter estimation was performed on 100 bootstrapped mean time series for each condition and observer. Each bootstrapped time series was formed by randomly resampling the underlying set of trials with replacement, with the number of trials per sample equal to the original number of trials. This produced 100-element distributions of each model parameter for each condition and observer. The median of a given parameter distribution was taken as the bootstrapped estimate of that parameter. The uncertainty of this estimate was quantified as the 95% confidence interval.

Parameter recovery To evaluate the parameter estimation procedure, we fit the model to artificial data generated by the model, with no noise. Because the form of the noise in pupil data is unknown, we relied on the bootstrapping procedure (in which the noise comes from the data itself) to quantify the uncertainty of parameter estimates and performed parameter recovery only to verify the accuracy of the fitting procedure and check for redundancies within the model structure. A set of 100 artificial time series was simulated by generating 100 sets of model parameters independently sampled from uniform distributions and calculating the resulting time series for each. Event and task-related amplitude values were sampled from a range of 0 to 10%, latency values were sampled from –500 to 500 ms, t_{max} values were sampled from 500 to 1500 ms, and y-int values were sampled from –4 to 4%. Response time values used in the boxcar function for the task-related component varied from 2350 to 3350 ms after the precue. The parameter estimation procedure was performed on each artificial time series (without bootstrapping), producing an output set of parameters for each input set of parameters. To evaluate the parameter recovery, input parameters were plotted against output parameters and the Pearson correlation coefficient was calculated.

Statistical testing

Hypothesis testing was performed using the median of the parameter estimates from the bootstrapping procedure. A linear mixed effects model was used to analyze the combined data across two experiments, each with a within-observer design, and in which three observers completed both experiments. A linear mixed effects model was created using the *lme4* package in R, with experiment and precue condition as fixed effects and observer as a random effect. We tested for main effects and interactions by approximating likelihood ratio tests to compare models with and without the effect of interest.

Results

We evaluated the ability of general linear models (GLMs) to capture pupil area time series during experimental trials with rapid sequences of events. We tested the model on a sample data set in which four sequential stimuli were presented within 2.25 s on each trial (Fig. 1a, b, see Methods). Given that 2 s is the approximate length of a typical PuRF (Hoeks & Levelt, 1993), we asked whether pupil responses to the successive events within a trial can be meaningfully recovered and what model form would best describe the pupil area time series over the course of a trial.

The data set included two experiments with the same stimulus sequence but different types of behavioral reports (orientation discrimination or orientation estimation) at the end of each trial. The experiments were previously reported with only the behavioral analysis (Denison et al., 2017) and with microsaccade analysis (Denison et al., 2019). Pupil area time series, averaged across trials for each experimental condition, were highly consistent within individual observers but varied considerably across observers (Fig. 1c), motivating an individual-level modeling approach.

The modeling framework assumed that the pupil response time series is a linear combination of component responses to various trial events, along with a task-related pupil response in each trial (Fig. 2) (de Gee et al., 2014; Hoeks & Levelt, 1993; Wierda et al., 2012). Each trial event was modeled as an impulse of variable amplitude (Fig. 2b), which was convolved with a PuRF (Fig. 2a) to generate the corresponding component response (Fig. 2c). The sum of all component responses was the predicted pupil time series (Fig. 2c).

Model comparison: Timing parameters improve fits

We compared 24 alternative models to determine what model structure would allow the best prediction of the pupil response time series. In particular, we asked whether the addition of two timing parameters would improve model fits over that of the standard model. The first timing parameter was event latency: a trial event impulse could have a non-zero latency with respect to its corresponding event, rather than being locked to the event onset. The second timing parameter was t_{max}: the time-to-peak of the PuRF could vary across individuals. We also tested different forms of the sustained, task-related pupil response, as well as the inclusion of a baseline parameter (y-int) to account for differences not removed by pre-trial baseline normalization. We used factorial model comparison (Keshvari, van den Berg, & Ma, 2012; Ma, 2018; van den Berg, Awh, & Ma, 2014) to test the contribution of each of these parameters to predicting pupil response time series (Table 1).

All four tested parameters (event latency, t_{max}, task-related component, and y-int) significantly improved model fits (Fig.

3; multi-way within-observers ANOVA on BIC scores, main effects of latency: $F(1,20) = 143.39$, $p = 1.4e\text{-}10$, mean across observers and models ΔBIC = –5,466; t_{max}: $F(1,20) = 76.59$, $p = 2.8e\text{-}08$, ΔBIC = –10,324; task-related: $F(2,40) = 19.88$, $p = 1.0e\text{-}06$, ΔBIC (box minus linear) = –1,379, ΔBIC (box minus none) = –8,558; *y*-int: $F(1,20) = 20.53$, $p = 2.0$ e-04, ΔBIC = –6,865). We also observed interactions between some factors, task-related component by t_{max}: $F(2,40) = 23.66$, $p = 1.7e\text{-}07$; t_{max} by *y*-int: $F(1,20) = 8.27$, $p = 9.4e\text{-}03$; task-related component by latency: $F(2,40) = 3.25$, $p = 4.9e\text{-}02$; t_{max} by latency: $F(1,20) = 7.00$, $p = 1.6e\text{-}02$. The best model for most observers (Model 1, best for 11 out of 21 observers) included all the tested parameters, with the task-related component modeled as a boxcar convolved with the PuRF. The model fit the mean time series data well (mean across observers, $R^2 = 0.99$; cross-validated $R^2 = 0.67$, 98% of noise ceiling). The single-trial R^2 was 0.20; this value reflects the noise at the single-trial level. Model 2 (best for six out of 21 observers), in which the task-related component was modeled as a linear function of time but which was otherwise identical to Model 1, performed similarly ($R^2 = 0.99$; cross-validated $R^2 = 0.68$, 98% of noise ceiling; single-trial $R^2 = 0.20$). These two types of task-related components also performed similarly at the factor level (ΔBIC = –1,379). Otherwise, Model 1 was significantly better than every other model (paired *t*-tests of Model 1 BIC vs. each model's BIC, all $t > 3.09$, all $p < 5.8e\text{-}03$ uncorrected; with Bonferroni correction for 23 pairwise comparisons, all but Model 3 had $p < 0.05$). Model 3, which was identical to Models 1 and 2 but with no task-related component, was the only other model that was best for some individuals, four out of 21 observers. We found a similar pattern across models for cross-validated R^2 as well as when we fit to single-trial data. Model 1 was consistently the best model, and latency, t_{max}, and task-related parameters improved model fits. Therefore, the addition of timing parameters to the standard model substantially improved model fits.

Our final test of the model's structure was to ask whether modeling all trial events was needed to predict the pupil area time series. In particular, the two visual target events were separated by only 250 ms, which is short compared to the dynamics of the pupil response. To test whether the model captured separate pupil responses to the two targets, we compared models that included either two target events (Model 1) or only one target event. The two-target model outperformed the one-target model, ΔBIC = –1,206, *t*(20)

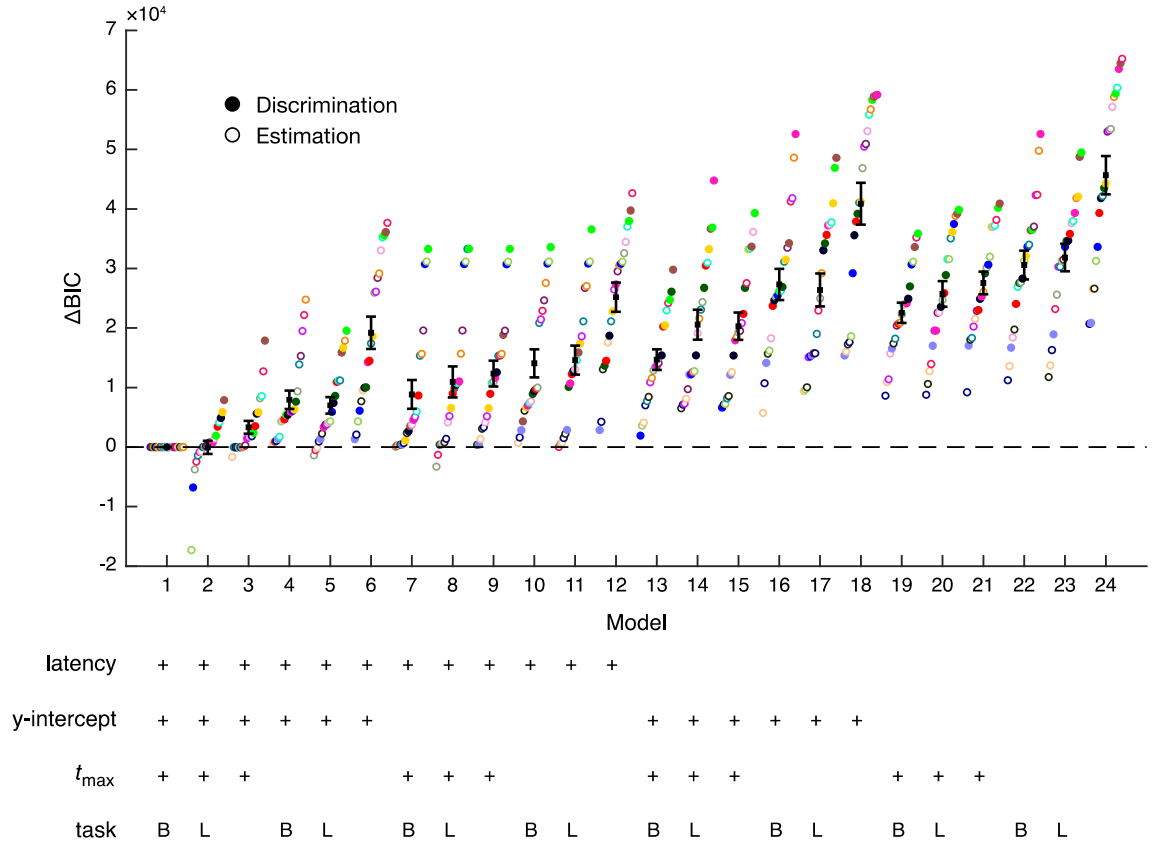

Fig. 3 Model comparison. Difference in BIC score for each observer and model with respect to Model 1, the best model for most observers. Each *color* corresponds to an individual observer. The *black square with error bar* corresponds to mean difference in BIC score across all observers. *Crosses* in the table indicate that the parameter type was fit for that model. In the task row, B indicates boxcar and L indicates linear function

= 4.23, p = 4.1e-04, consistent with separate pupil responses to the two rapidly presented targets.

Validation of the fitting procedure: Parameter recovery and tradeoffs

We evaluated the best model (Model 1) and fitting procedures in several ways. First, we sought to validate the model and fitting procedures by performing parameter recovery on simulated data. Redundancies in the model or lack of precision to resolve the unique contributions of different trial events to the pupil time series would result in parameter tradeoffs, and noise in the fitting procedure from stochastically searching a high-dimensional parameter space would result in variability in the parameter estimates. We performed parameter recovery to assess these possibilities by generating 100 simulated time series with known parameter values and then fitting the model. Parameter estimates from simulated data tended to be similar to the true values for the entire range of tested values (strong diagonals on the 2D histograms in Fig. 4 and correlations in each panel). This was also the case when the range of T1-T2 SOAs was restricted to ± 100 ms around the experimental SOA of 250 ms (Fig. S1), as well as when noise on single trials was simulated (Fig. S2). Parameter recovery accuracy was lowest for the T1 and T2 amplitudes, likely because of their close temporal proximity. Figure S1 indicates the level of recovery precision that can be expected for the shortest SOA range tested (r = 0.52–0.54 vs. r = 0.60–0.69 for all SOAs). Accuracy for the timing parameters was generally high.

We assessed parameter tradeoffs first by examining correlations between estimated values for pairs of parameters in the simulated data. No correlations were significant after Bonferroni correction for multiple comparisons (Fig. S3). The lack of significant correlations between parameters in the simulated data indicates that parameter tradeoffs are not inherent to the structure of the model or the fitting procedure.

We next assessed parameter tradeoffs by examining correlations between pairs of parameter values (bootstrap medians) estimated from the real data. All event-related amplitudes were positively correlated with each other (r > 0.61, p < 1.2e-07). Certain event latencies were also positively correlated with each other (T1 with T2, T2 with response cue; r > 0.44, p < 2.7e-04). These correlations are likely to arise from true statistical dependencies in the data; e.g., some observers have generally stronger pupil dilation responses, across all events. In addition to these positive correlations, T1 latency negatively correlated with event-related amplitudes (precue, T1, T2; r < –0.44, p < 3e-04) and precue latency negatively correlated with task-related amplitude (r = –0.48, p < 6.1e-05). t_{max} negatively correlated with y-int (r = –0.43, p = 4.7e-04) and positively correlated with response cue latency (r = 0.53, p = 6.4e-06). The negative correlations could arise from parameter tradeoffs driven by noise in the real data, or from true statistical dependencies in the pupil responses.

Reliability of parameter estimates for individual observers

We next evaluated the reliability of the parameter estimates in real data. Real data have multiple sources of noise, some of which are unknown, so to estimate the reliability of parameter estimates given such noise, we used a bootstrapping procedure. This procedure allowed us to estimate the reliability of parameter estimates for individual observers. Parameter estimates and their reliabilities for an example observer are shown in Fig. 5. We define "reliability" as the range of the 95% confidence interval. The reliability for each parameter

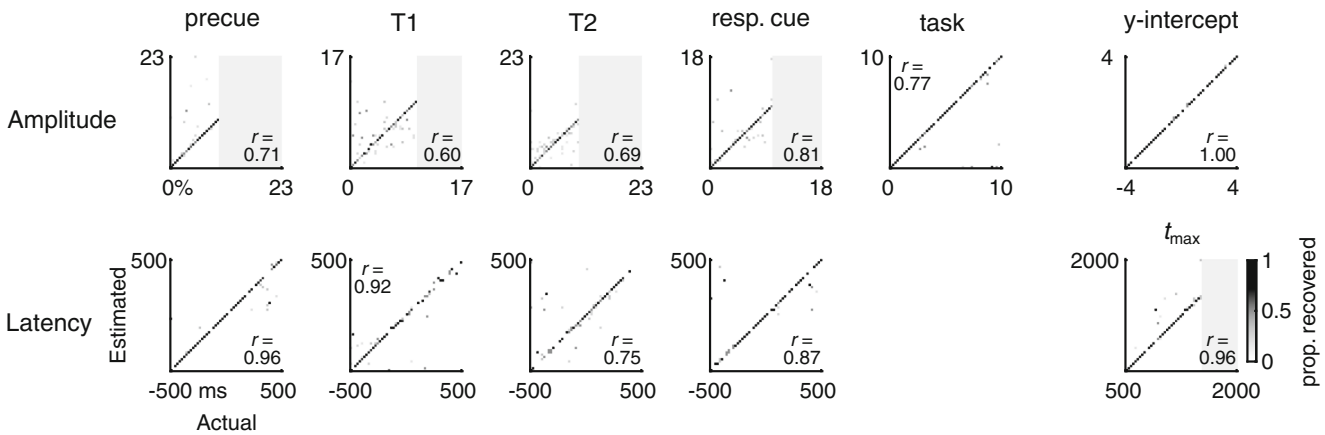

Fig. 4 Parameter recovery for simulated data. 2D histograms showing the proportion of model fits for which a given actual parameter value (x-axis) was fit as a given estimated parameter value (y-axis). Perfect model recovery would appear as a diagonal black line (all actual values recovered as the same estimated values). Each panel is one parameter. Pearson correlations between actual and estimated values are given in each panel. *Gray shaded regions* are outside the range of actual parameter values simulated

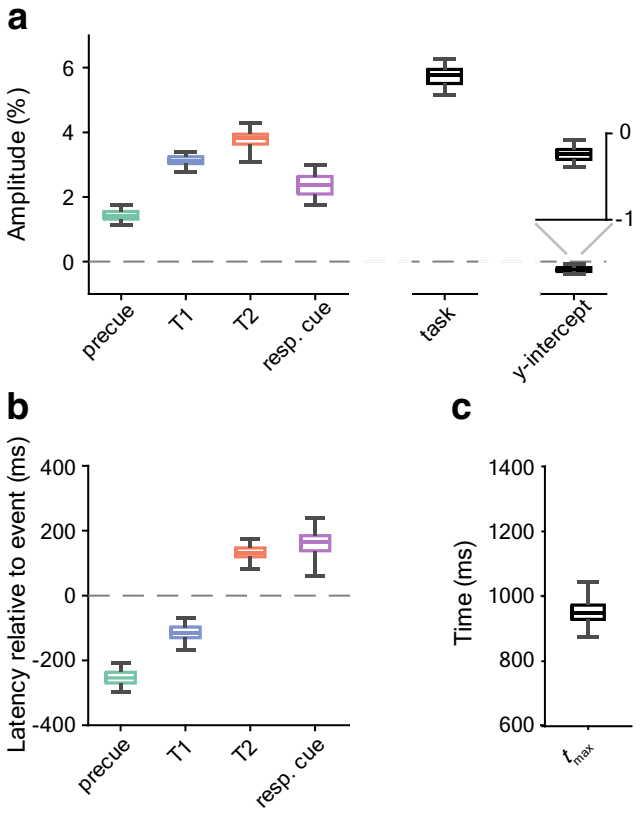

Fig. 5 Reliability of individual parameter estimates for a representative observer. **a** Amplitude and *y*-int estimates. *Box and whisker plots* show bootstrap median along with 50% and 95% bootstrapped confidence intervals. **b** Latency and **c** t_{max} estimates, plotted as in panel **a**

estimate from the example observer is shown in Fig. 5. The mean reliabilities of different event types across observers were: trial event amplitude: 2.21%; task-related amplitude: 2.75%; trial event latency: 342 ms; t_{max}: 334 ms; *y*-int: 0.74%. Due to this variability across bootstrap samples, we recommend using the median of the bootstrapped distribution as a robust parameter estimate, and we have adopted this practice here.

Consistency of parameter estimates within observers and variability across observers

We next investigated the consistency of parameter estimates within each observer as well as their variability across observers. To do this, we measured the consistency of parameter estimates for a given observer across independent sets of trials (Fig. 6). We split the trials for each observer based on the experimental condition (three conditions per observer: precue T1, T2, neutral). Parameter estimates were generally consistent for individual observers, though latency was less consistent than the other parameters. In contrast to the within-observer consistency, all parameters varied substantially across observers. In particular, the t_{max} parameter, which describes the dynamics of the PuRF, was highly consistent within individual observers but varied over a large range across observers (from ~700 to ~1600 ms). These findings underscore the importance of modeling individual pupil response time series rather than only considering a group average time series, and they further show the importance of modeling individual observer pupil response dynamics rather than assuming a fixed PuRF. Averaging observers' time series would blur the distinct individual dynamics; using a single PuRF for all observers would result in misleading parameter estimates due to model mismatch.

Parameter estimates for the temporal attention experiment: Amplitude depends on task

To demonstrate how modeling can reveal cognitive modulations of the pupil response, we used the developed modeling procedure to evaluate the parameter estimates in the experimental data. We calculated separate estimates for each precue type (T1, T2, neutral), separately for the Discrimination and Estimation experiments, which had the same stimulus sequence but different types of behavioral reports (Fig. 7). No differences were found among the different precue types ($\chi^2(2) < 5.77$, $p > 0.05$ for all parameters). There were also no interactions between precue and experiment ($\chi^2(2) < 3.70$, $p > 0.15$). So here we report the mean across precues for each experiment.

Amplitude estimates for trial events were 1–7% change from baseline, and the amplitude of the decision-related signal was 2.18% (Discrimination) or 3.44% (Estimation). The mean *y*-int was slightly below zero, driven by a few observers with larger negative *y*-int values (Fig. 6), and did not differ between experiments (–0.50% for Discrimination, –0.54% for Estimation). Interestingly, amplitude estimates were higher for all trial events in the Discrimination experiment compared to the Estimation experiment (Fig. 7a, $\chi^2(1) > 32.62$, $p < 1.3e\text{-}7$ with Bonferroni correction for multiple comparisons across parameters). To assess whether this effect was present at an individual observer level, we examined the trial event amplitudes of the three observers who participated in both experiments. We found that two out of the three observers, like the group data, had higher event amplitudes for Discrimination compared to Estimation (differences of 6.20% and 8.49%), whereas one had similar amplitudes (difference of –0.24%). No other task differences survived correction for multiple comparisons.

Latency estimates were similar for the two experiments (Fig. 7b, $\chi^2(1) < 1.75$, $p > 0.18$). The latency estimate for T2 was similar to the event onset, (51 ms for Discrimination, 19 ms for Estimation, comparison to zero latency: $\chi^2(1) = 2.62$, $p = 0.11$). However, the latency estimate for T1 was well before T1 onset (–290 ms for Discrimination, –182 ms for

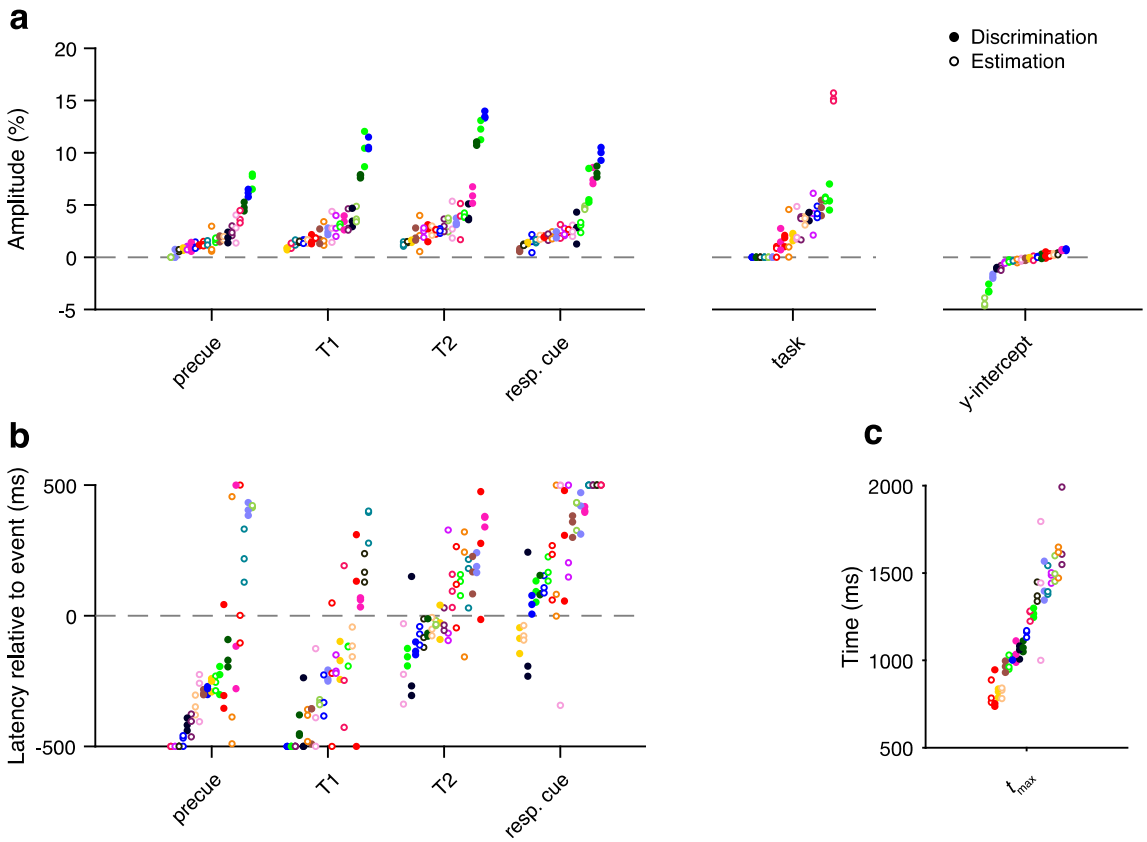

Fig. 6 Consistency of parameter estimates across independent sets of trials. **a** Amplitude and *y*-int estimates. Each point is one condition (precue T1, T2, neutral) for one observer; each condition was fit separately. Each observer has a unique color. *Filled points* are from the Discrimination experiment and *empty points* are from the Estimation experiment. **b** Latency and **c** t_{max} estimates, plotted as in panel **a**

Estimation, comparison to zero: $\chi^2(1) = 65.88$, $p = 1.9e\text{-}15$ corrected). The precue latency estimate was also well before the precue onset (–157 ms for Discrimination, –221 ms for Estimation, comparison to zero: $\chi^2(1) = 32.13$, $p = 5.8e\text{-}8$ corrected). Meanwhile, the response cue latency estimate was delayed relative to the response cue onset (169 ms for Discrimination, 292 ms for Estimation, comparison to zero: $\chi^2(1) = 89.84$, $p < 8e\text{-}16$ corrected). The mean t_{max} was 1,053 ms for Discrimination and 1,296 for Estimation, with no significant difference ($\chi^2(1) = 0.12$, $p = 0.73$). Thus, the two experiments had similar latency but different amplitude profiles, with larger event-related pupil responses in the Discrimination than the Estimation experiment.

Discussion

Pupil size is an accessible, continuous measure that reflects rapidly changing internal states, but the pupil response itself is relatively slow. Linear modeling has shown promise for inferring the dynamics of internal signals that drive pupil responses, but as has been noted (Bach et al., 2018), these methods have not been systematically evaluated. To work toward a standard pupil modeling approach, here we compared different pupil models, validated modeling procedures, and evaluated the reliability of the best model. Based on the results, we recommend a specific pupil model and fitting procedure, and we quantify the uncertainty of the resulting parameter estimates. The best model includes timing parameters that are not usually fit, indicating that more precise modeling of pupil dynamics may improve the estimation of pupillary responses to rapid events.

Model validation

Despite the increasing use of linear models of pupil size to capture pupillary time series to multiple sequential events, such methods have not been well validated. Our best model and fitting procedures performed well on simulated data, with reasonably accurate parameter recovery. The model also fit the real data well. The unknown nature of noise in the pupil data limited our ability to simulate the impact of noise on parameter recovery, so we also quantified the uncertainty of the parameter estimates in the real data. Note that these uncertainty estimates are expected to depend on the number of trials. We also identified a few parameter tradeoffs in the real data. Both uncertainty and tradeoffs should be considered when

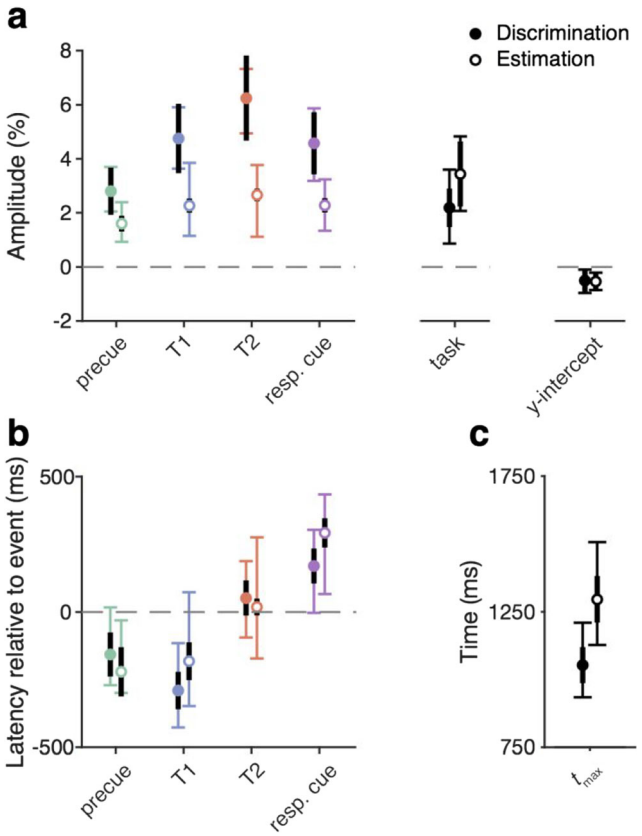

Fig. 7 Group parameter estimates. **a** Amplitude and *y*-int estimates. *Colored points* show mean of bootstrap medians across observers and conditions. *Filled points* are from the Discrimination experiment, *empty points* are from the Estimation experiment. Thin *colored error bars* show mean 95% confidence intervals across observers. *Thick black error bars* show SEM across observers. **b** Latency and **c** t_{max} estimates, plotted as in panel **a**

interpreting parameter estimates and potentially when designing experiments. For example, given the ~350 ms 95% confidence interval on latency estimates, it may be helpful to separate successive trial events by at least that interval, if possible. The results suggest that the current model is reasonable as a current standard and can serve as a starting point for future work.

Temporal properties

The inclusion of two timing parameters, event latency and t_{max}, improved the model's ability to fit the pupil size time series. With respect to latency, internal signals related to the precue and T1 events were estimated to occur before the events themselves. This finding suggests that these internal signals anticipated the stimulus onsets, which were predictable. Pupil dilation in advance of a predictable stimulus has been observed previously and found to depend on temporal expectation (Akdoğan, Balcı, & van Rijn, 2016; Bradshaw, 1968) and upcoming task demands (Irons, Jeon, & Leber, 2017). Allowing for variable latency in pupil models may therefore be particularly important when observers have expectations about the timing of upcoming events. More broadly, latency estimates can provide information about anticipatory processes related to the observer's task.

Most previous pupil modeling studies (de Gee et al., 2017; de Gee et al., 2014; Kang et al., 2014; Kang & Wheatley, 2015; Lempert et al., 2015; Murphy et al., 2016; van Kempen et al., 2019; Wierda et al., 2012; Willems, Damsma, et al., 2015; Willems, Herdzin, & Martens, 2015; Zylberberg et al., 2012) have used the canonical PuRF proposed by Hoeks and Levelt (1993), which assumes that all observers have identical pupil dynamics. We found, on the contrary, that fitting the time-to-peak (t_{max}) of the PuRF improved model fits. The value of t_{max} varied widely across observers but was highly consistent for a given observer, suggesting that t_{max} is an observer-specific property. The t_{max} values we estimated for individual observers were in line with Hoeks and Levelt's original estimates using a single stimulus event, which ranged from 630–1300 ms and showed some variability between auditory vs. visual events (Hoeks & Levelt, 1993). van den Brink et al. (2016) also varied t_{max}, but did so by setting it to the latency of the maximum dilation in the time series, rather than fitting it. Here we found that individual PuRFs can be estimated from a multi-event time series and should be used instead of the canonical PuRF to improve pupil modeling.

Despite the sluggishness of the pupillary response, a model with two target events outperformed a model with only one target event. This suggests that the two target events were associated with separate pupil dilations, even though they were separated by only 250 ms. Due to the early response to T1, however, the estimated dilations occurred further apart in time, closer to 500 ms. Individual observer latency estimates had a reliability of ~350 ms, indicating that while separate pupillary responses to events close in time seem to be recoverable, one should take care in interpreting their exact timing. The interpretation of some estimated latencies in the current data set was also limited by the fact that they fell at the boundary of the allowed range, –500 ms.

Dependence on task and temporal attention

The event-related pupil response amplitude depended on the task observers were performing, with larger amplitudes in the Discrimination compared to the Estimation task. This finding demonstrates that even when the stimulus sequence is identical, cognitive factors can influence the pupil response, consistent with a large body of research (Ebitz & Moore, 2018; Einhäuser, 2017; Mathôt, 2018). Here, a relatively modest change in task instruction—discrimination vs. estimation—changed the amplitudes of pupillary responses to sensory stimuli. The larger amplitude effect for Discrimination could have been due, at least in part, to a larger baseline pupil size in

the Estimation task, perhaps related to a higher cognitive load, as tonic size and phasic response amplitudes are inversely related (Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994; de Gee et al., 2014; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Murphy, Robertson, Balsters, O'Connell, & G., 2011). We were unable to directly compare baseline pupil size across tasks, however, because the tasks were performed in separate sessions. Task affected pupil response amplitudes to trial events more than the sustained, task-related amplitude, and had no effect on pupil response latencies. These results show how models can help specify the effects of cognitive manipulations on pupillary responses.

In contrast, we found no reliable impact of temporal attention on any model parameter, despite finding overall effects of temporal expectation in the form of anticipatory responses to the predictably timed stimuli, as well as behavioral effects of temporal attention in the same experiments (Denison et al., 2017). While it is difficult to draw any strong conclusions from a null result, our findings suggest that the effects of voluntary temporal attention on pupil size may be subtle, if not absent. Previous research has linked changes in pupil size to temporal selection during the attentional blink, in which observers identify targets embedded in a rapid visual sequence (Wierda et al., 2012; Willems, Damsma, et al., 2015; Willems, Herdzin, & Martens, 2015; Zylberberg et al., 2012). One model of the attentional blink explains the phenomenon as arising from the dynamics of the LC (Nieuwenhuis, Gilzenrat, Holmes, & Cohen, 2005), according to the idea that phasic LC responses act as a temporal filter (Aston-Jones & Cohen, 2005). Temporal selection during the attentional blink—in which target timing is unpredictable—may be different, however, from voluntary temporal attention—the prioritization of specific, relevant time points that are fully predictable. In contrast to our findings for the pupil, microsaccade dynamics are modulated by both temporal expectation (Amit, Abeles, Carrasco, & Yuval-Greenberg, 2019; Dankner, Shalev, Carrasco, & Yuval-Greenberg, 2017; Denison et al., 2019; Hafed, Lovejoy, & Krauzlis, 2011; Pastukhov & Braun, 2010) and temporal attention (Denison et al., 2019).

Limitations and extensions

The best model was determined based on a sample dataset in which ~ 2-s trials contained multiple sequential events. This dataset was therefore suited to ask about the estimation of rapid internal signals driving pupil dilation. Nevertheless, other models may perform better for different datasets. For example, Murphy et al., 2016, showed that the task-related component, which we found to be best modeled as a boxcar, was better described as boxcar or linear depending on the task. van Kempen et al., 2019, also found support for a linear task-related component. Note that all of these best-fitting boxcar and linear regressors were dependent on RT, unlike the linear regressor we tested, which modeled linear drift across the whole trial following Wierda et al., 2012.

Future work could extend the current model in multiple ways. In the main analyses, we modeled average pupil time series across trials in a given experimental condition, using the median RT to define the task-related boxcar component. Another approach would be to model single trial time series and define the boxcar for each trial using that trial's RT (e.g., de Gee et al., 2014). In the present study, we tested only the t_{max} parameter of the PuRF, leaving the second parameter, n, fixed. The motivations for this choice were that (1) t_{max} is easily interpretable as the time-to-peak of the PuRF, whereas n governs the shape of the PuRF in a more complex way, and (2) Hoeks and Levelt reported that n could vary considerably without a large impact on the other parameter estimates. However, future work could test whether fitting n would further improve the model fits. Future work could also add covariates to the model such as tonic pupil size (de Gee et al., 2014), model pupil modulations associated with blinks and microsaccades (Knapen et al., 2016), and model pupil constrictions as well as dilations (Korn & Bach, 2016). A recent pupil model included both transient and sustained components of the pupil dilation response (Spitschan et al., 2017), which could be compared to the unitary PuRF used here and in most previous work. Furthermore, it will be important to test this model on a variety of perceptual and cognitive tasks, including those known to modulate pupil responses (Einhäuser, 2017). The PRET toolbox will facilitate tests to address generalization of the present study to multiple tasks and observer populations.

General linear models have also been used to model BOLD fMRI time series. A common strategy for accommodating individual variations in hemodynamic response function (HRF) timing has been to use temporal derivatives of the canonical HRF to generate extra regressors in the model. Such an approach could also be used to model the PuRF, with the practical advantage that the models could then be fit using regression, which is less computationally demanding than the optimization procedure used here. We chose to parameterize the latency and t_{max} of the PuRF instead of introducing separate temporal derivative regressors for two reasons. The first is interpretability. We can read out the latency and t_{max} parameter estimates directly from our model, whereas they would have to be derived if multiple PuRF regressors were used. The second reason is theoretical: Latency and t_{max} have different underlying sources so they should be decoupled in the model. Specifically, t_{max} is due to pupillary mechanics, whereas latency also depends on the latency of the neural response that drives the pupil dilation. Letting these parameters be independent therefore allowed us to fit a single t_{max} per observer but separate latency parameters for each event, better reflecting their different underlying sources.

Pupil Response Estimation Toolbox (PRET)

We provide a MATLAB toolbox, PRET, that performs all the analyses reported here, including basic preprocessing (blink interpolation and baseline correction), model specification, model fitting, bootstrapping of parameter estimates, data simulation and parameter recovery, and model comparison. The toolbox is open-source and freely available on GitHub (https://github.com/jacobaparker/PRET). Importantly, PRET can be readily employed for model comparison and uncertainty estimation in other data sets to continue to work toward a field-standard pupil modeling approach.

Open Practices Statement

The code for the pupil analyses is available at https://github.com/jacobaparker/PRET. The data for all experiments are available upon request. None of the experiments was preregistered.

Author Notes This research was supported by National Institutes of Health National Eye Institute R01 EY019693 and R01 EY016200 to M.C., F32 EY025533 to R.N.D., and T32 EY007136 to NYU. R.N.D., J.A.P., and M.C. designed research; R.N.D. collected data; R.N.D. and J.A.P. performed analyses and wrote the manuscript; M.C. edited the manuscript; J.A.P. authored the PRET toolbox under the supervision of R.N.D.

Compliance with ethical standards

Conflicts of interest The authors have no conflicts of interest to declare.

References

Akdoğan, B., Balcı, F., & van Rijn, H. (2016). Temporal expectation indexed by pupillary response. *Timing & Time Perception, 4*(4), 354–370. https://doi.org/10.1163/22134468-00002075

Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *eLife, 5*, 403. https://doi.org/10.7554/eLife.18103

Amit, R., Abeles, D., Carrasco, M., & Yuval-Greenberg, S. (2019). Oculomotor inhibition reflects temporal expectations. *NeuroImage, 184*, 279–292. https://doi.org/10.1016/j.neuroimage.2018.09.026

Aston-Jones, G., & Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *Journal of Comparative Neurology, 493*(1), 99–110. https://doi.org/10.1002/cne.20723

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience, 14*(7), 4467–4480.

Bach, D. R., Castegnetti, G., Korn, C. W., Gerster, S., Melinscak, F., & Moser, T. (2018). Psychophysiological modeling: Current state and future directions. *Psychophysiology, 55*(11), e13214. https://doi.org/10.1111/psyp.13209

Binda, P., & Gamlin, P. D. (2017). Renewed attention on the pupil light reflex. *Trends in Neurosciences, 40*(8), 455–457. https://doi.org/10.1016/j.tins.2017.06.007

Bradshaw, J. L. (1968). Pupillary changes and reaction time with varied stimulus uncertainty. *Psychonomic Science, 13*(2), 69–70. https://doi.org/10.3758/BF03342414

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Castañón, S. H., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron, 81*(6), 1429–1441. https://doi.org/10.1016/j.neuron.2014.01.020

Dankner, Y., Shalev, L., Carrasco, M., & Yuval-Greenberg, S. (2017). Prestimulus inhibition of saccades in adults with and without attention-deficit/hyperactivity disorder as an index of temporal expectations. *Psychological Science, 28*(7), 835–850. https://doi.org/10.1177/0956797617694863

de Gee, J. W., Colizoli, O., Kloosterman, N. A., Knapen, T., Nieuwenhuis, S., & Donner, T. H. (2017). Dynamic modulation of decision biases by brainstem arousal systems. *eLife, 6*, 309. https://doi.org/10.7554/eLife.23232

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences, 111*(5), E618-625. https://doi.org/10.1073/pnas.1317557111

Denison, R. N., Heeger, D. J., & Carrasco, M. (2017). Attention flexibly trades off across points in time. *Psychonomic Bulletin & Review, 24*(4), 1142–1151. https://doi.org/10.3758/s13423-016-1216-1

Denison, R. N., Yuval-Greenberg, S., & Carrasco, M. (2019). Directing voluntary temporal attention increases fixational stability. *Journal of Neuroscience, 39*(2), 353–363. https://doi.org/10.1523/JNEUROSCI.1926-18.2018

Ebitz, R. B., & Moore, T. (2018). Both a gauge and a filter: Cognitive modulations of pupil size. *Frontiers in Neurology, 9*, 1190. https://doi.org/10.3389/fneur.2018.01190

Einhäuser, W. (2017). The pupil as marker of cognitive processes. In Q. Zhao (Ed.), *Computational and Cognitive Neuroscience of Vision* (pp. 141–169). Singapore: Springer Singapore.

Fernandez, A., Denison, R. N., & Carrasco, M. (2019). Temporal attention improves perception similarly at foveal and parafoveal locations. *Journal of Vision, 19*(1), 12. https://doi.org/10.1167/19.1.12

Foote, S. L., Aston-Jones, G., & Bloom, F. E. (1980). Impulse activity of locus coeruleus neurons in awake rats and monkeys is a function of sensory stimulation and arousal. *Proceedings of the National Academy of Sciences, 77*(5), 3033–3037.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Hafed, Z. M., Lovejoy, L. P., & Krauzlis, R. J. (2011). Modulation of microsaccades in monkey during a covert visual attention task. *Journal of Neuroscience, 31*(43), 15219–15230. https://doi.org/10.1523/JNEUROSCI.3106-11.2011

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349–350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers, 25*(1), 16–26. https://doi.org/10.3758/BF03204445

Irons, J. L., Jeon, M., & Leber, A. B. (2017). Pre-stimulus pupil dilation and the preparatory control of attention. *PLoS One, 12*(12), e0188787. https://doi.org/10.1371/journal.pone.0188787

Johansson, B., & Balkenius, C. (2017). A computational model of pupil dilation. *Connection Science, 30*(1), 5–19. https://doi.org/10.1080/09540091.2016.1271401

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron, 89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583–1585.

Kahneman, D., Beatty, J., & Pollack, I. (1967). Perceptual deficit during a mental task. *Science, 157*(3785), 218–219. https://doi.org/10.1126/science.157.3785.218

Kang, O. E., Huffer, K. E., & Wheatley, T. P. (2014). Pupil dilation dynamics track attention to high-level information. *PLoS One, 9*(8), e102463. https://doi.org/10.1371/journal.pone.0102463

Kang, O. E., & Wheatley, T. (2015). Pupil dilation patterns reflect the contents of consciousness. *Consciousness and Cognition, 35*, 128–135. https://doi.org/10.1016/j.concog.2015.05.001

Keshvari, S., van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS One, 7*(6), e40216. https://doi.org/10.1371/journal.pone.0040216

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? ECVP Abstract Supplement *Perception, 36*.

Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A. F., Bonneh, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience, 41*(8), 1068–1078. https://doi.org/10.1111/ejn.12859

Knapen, T., de Gee, J. W., Brascamp, J., Nuiten, S., Hoppenbrouwers, S., & Theeuwes, J. (2016). Cognitive and ocular factors jointly determine pupil responses under equiluminance. *PLoS One, 11*(5), e0155574. https://doi.org/10.1371/journal.pone.0155574

Korn, C. W., & Bach, D. R. (2016). A solid frame for the window on cognition: Modeling event-related pupil responses. *Journal of Vision, 16*(3), 28, 1–16. https://doi.org/10.1167/16.3.28

Korn, C. W., Staib, M., Tzovara, A., Castegnetti, G., & Bach, D. R. (2017). A pupil size response model to assess fear learning. *Psychophysiology, 54*(3), 330–343. https://doi.org/10.1111/psyp.12801

Lempert, K. M., Chen, Y. L., & Fleming, S. M. (2015). Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS One, 10*(5), e0126588. https://doi.org/10.1371/journal.pone.0126588

Libby, W. L., Lacey, B. C., & Lacey, J. I. (1973). Pupillary and cardiac activity during visual attention. *Psychophysiology, 10*(3), 270–294.

Ma, W. J. (2018). Identifying suboptimalities with factorial model comparison. *Behavioral and Brain Sciences, 41*, e234. https://doi.org/10.1017/S0140525X18001541

Mathôt, S. (2013). A simple way to reconstruct pupil size during eye blinks. https://doi.org/10.6084/m9.figshare.688001

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition, 1*(1):16, 1–23. https://doi.org/10.5334/joc.18

Mathôt, S., Fabius, J., Heusden, E., & Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods, 50*, 94–106. https://doi.org/10.3758/s13428-017-1007-2

Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications, 7*(1), 13526. https://doi.org/10.1038/ncomms13526

Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping, 35*(8), 4140–4154. https://doi.org/10.1002/hbm.22466

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'Connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. *Psychophysiology, 48*(11), 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x

Murphy, P. R., Vandekerckhove, J., & Nieuwenhuis, S. (2014). Pupil-linked arousal determines variability in perceptual decision making. *PLoS Computational Biology, 10*(9), e1003854. https://doi.org/10.1371/journal.pcbi.1003854

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience, 15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Nieuwenhuis, S., Gilzenrat, M. S., Holmes, B. D., & Cohen, J. D. (2005). The role of the locus coeruleus in mediating the attentional blink: A neurocomputational theory. *Journal of Experimental Psychology: General, 134*(3), 291–307. https://doi.org/10.1037/0096-3445.134.3.291

Nobre, A. C., & van Ede, F. (2018). Anticipated moments: Temporal structure in attention. *Nature Reviews Neuroscience, 19*(1), 34–48. https://doi.org/10.1038/nrn.2017.141

Pastukhov, A., & Braun, J. (2010). Rare but precious: Microsaccades are highly informative about attentional allocation. *Vision Research, 50*(12), 1173–1184. https://doi.org/10.1016/j.visres.2010.04.007

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience, 5*: 115, 1–12. https://doi.org/10.3389/fnins.2011.00115

Reilly, J., Kelly, A., Kim, S. H., Jett, S., & Zuckerman, B. (2018). The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry. *Behavior Research Methods, 28*(1), 403–414. https://doi.org/10.3758/s13428-018-1134-4

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications, 7*, 13289. https://doi.org/10.1038/ncomms13289

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron, 76*(1), 130–141. https://doi.org/10.1016/j.neuron.2012.09.011

Spitschan, M., Bock, A. S., Ryan, J., Frazzetta, G., Brainard, D. H., & Aguirre, G. K. (2017). The human visual cortex response to melanopsin-directed stimulation is accompanied by a distinct perceptual experience. *Proceedings of the National Academy of Sciences, 114*(46), 12291–12296. https://doi.org/10.1073/pnas.1711522114

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications, 8*, 14637. https://doi.org/10.1038/ncomms14637

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124–149. https://doi.org/10.1037/a0035234

van den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PLoS One, 11*(10), e0165274.https://doi.org/10.1371/journal.pone.0165274

van Kempen, J., Loughnane, G. M., Newman, D. P., Kelly, S. P., Thiele, A., O'Connell, R. G., & Bellgrove, M. A. (2019). Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *eLife, 8*:e42541. https://doi.org/10.7554/eLife.42541

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *Journal of Neuroscience, 35*(20), 7866–7877. https://doi.org/10.1523/JNEUROSCI.0454-15.2015

Wang, C.-A., & Munoz, D. P. (2015). A circuit for pupil orienting responses: Implications for cognitive modulation of pupil size. *Current Opinion in Neurobiology, 33*, 134–140. https://doi.org/10.1016/j.conb.2015.03.018

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences, 109*(22), 8456–8460. https://doi.org/10.1073/pnas.1201858109

Willems, C., Damsma, A., Wierda, S. M., Taatgen, N., & Martens, S. (2015). Training-induced changes in the dynamics of attention as reflected in pupil dilation. *Journal of Cognitive Neuroscience, 27*(6), 1161–1171. https://doi.org/10.1162/jocn_a_00767

Willems, C., Herdzin, J., & Martens, S. (2015). Individual differences in temporal selective attention as reflected in pupil dilation. *PLoS One, 10*(12), e0145056. https://doi.org/10.1371/journal.pone.0145056

Zylberberg, A., Oliva, M., & Sigman, M. (2012). Pupil dilation: A fingerprint of temporal selection during the "attentional blink". *Frontiers in Psychology, 3*, 316. https://doi.org/10.3389/fpsyg.2012.00316