**Testing the model on additional data sets**

Our model is general and works with data collected using various task protocols, including cases in which the inter-stimulus interval duration is jittered across trials, and when task conditions are blocked or interleaved across trials. To illustrate this, we collected behavioral and eye data from three additional tasks and fit our model to these data.

The first task was identical to the hard runs we ran in our main experiment, with two modifications: (1) instead of a staircase controlling the stimulus tilt, we found each observer's 75% tilt threshold before the experiment and fixed the tilt at that value for the whole experiment; (2) the trials were 3, instead of 4, seconds long. We had five new observers perform this task. Consent was obtained from each and the same statements made in the "Participants" section of the Methods apply here and throughout to all other additional observers. In this task, there was no uncertainty about the stimulus tilt, i.e., external noise. The model fit each observer's data well (Fig. S15).

The second task was identical to the hard task from our main experiment, but a few key differences: (1) the stimulus was a grating in an annulus (outer diameter, 1.5 deg) surrounding the fixation cross. (2) The tilt of the grating on each trial was drawn from a Gaussian distribution whose standard deviation was determined by a thresholding procedure for each observer separately to achieve approximately 75% discrimination accuracy on average. (3) The task structure consisted of a 900 ms ISI, a 200 ms stimulus presentation, and a 900 ms response window. Thus, in this task, the stimulus was not in the periphery, but rather within the fovea surrounding fixation, meaning that saccades shouldn't be expected to occur more often in any particular direction. And task difficulty was interleaved across trials, and there was no cue indicating the difficulty of the task within a run. Fitting the model to this data from four new observers, we found that the model fit well, gain was modulated by difficulty under these new conditions, being larger for harder trials, and saccade rate was inhibited at stimulus onset in a similar way to during our main experiment (Fig. S16).

The third task was a block alternation task, identical to our main experiment, but with the key differences that the task difficulty changed every 5 trials, and that the trials were 2 seconds long. The fixation cross still changed color every time difficulty switched. The data, collected from a mix of four new and old observers, were fit well by the model. That said, one observer (O11) had a number of runs that were not fit as well and it's unclear why this occurred. The model fit the data from both the hard and easy blocks of trials, and gain was modulated by task difficulty, being higher for hard trials (Fig. S17). This demonstrates that the model can estimate arousal even for more complex task designs, in which arousal is changing at a faster rate.

To get at the question of whether our model can fit data from a task with jittered ISIs, we performed a simulation in which we used our forward model to generate synthetic pupil and saccade data from a version of our task in which the ISI duration on each trial (75 per run) was drawn from a uniform distribution between 2.4 to 4 s (i.e., jittered), and arousal level was alternated between runs (by setting the ground truth gain to 1 or 2, for easy and hard runs respectively). I.i.d. Gaussian noise was added to the generator function on each trial, so the pupil responses varied from trial to trial. We used our method to estimate gain and to predict pupil size from the saccade rate function (N simulated trials, 75). The average $R^2$ of the model fit was 99% and the average ratio of the estimated to ground truth gain modulation was 1.0005, i.e., perfect recovery. This shows that our model can fit well when the trial duration is variable. Note that we used the known pupil IRF for this simulation because the purpose was to show that our algorithm works for jittered timing, not to explore estimating the IRF. Of course the R2 will go down if the IRF is unknown and needs to be estimated as well (in that case, the $R^2$ was 85%, but the gain modulation recovery ratio was still 1, i.e., perfect). These results suggest that our model should likely generalize to tasks in which trial duration is variable.

**Variations and extensions**

One extension of our model is to estimate multiple gains per trial — for example one gain value before the button press and another gain value after. This is feasible as long as there is trial-to-trial variability in the behavior that could constrain this estimation.

Another useful extension of our model is to simultaneously estimate the components of the generator function that are locked to different task events, assuming a linear system. This would allow for better prediction of the task-evoked pupil response across trials because it can account for trial-to-trial variability in the generator function (e.g., variability generated by variability in reaction times). For example, we showed in Fig. S8 and S9 the generator function locked to trial onset and the button press, averaged across trials. Here, instead one would deconvolve the saccade rate function to the trial onset and the button press with one large block design matrix, and then estimate the two generator functions. The component locked to the button press could then be shifted on different trials to the times of button presses and summed with the repeated trial-onset-locked generator function. Then both could be scaled by the estimated gain or gains (if one is estimating multiple gains within a trial, for instance one before the button press and one after). This "generator time series" across trials would represent the input to the pupil across multiple trials. One could use this same approach to with other trial events as well, such as pre- and post-cues.

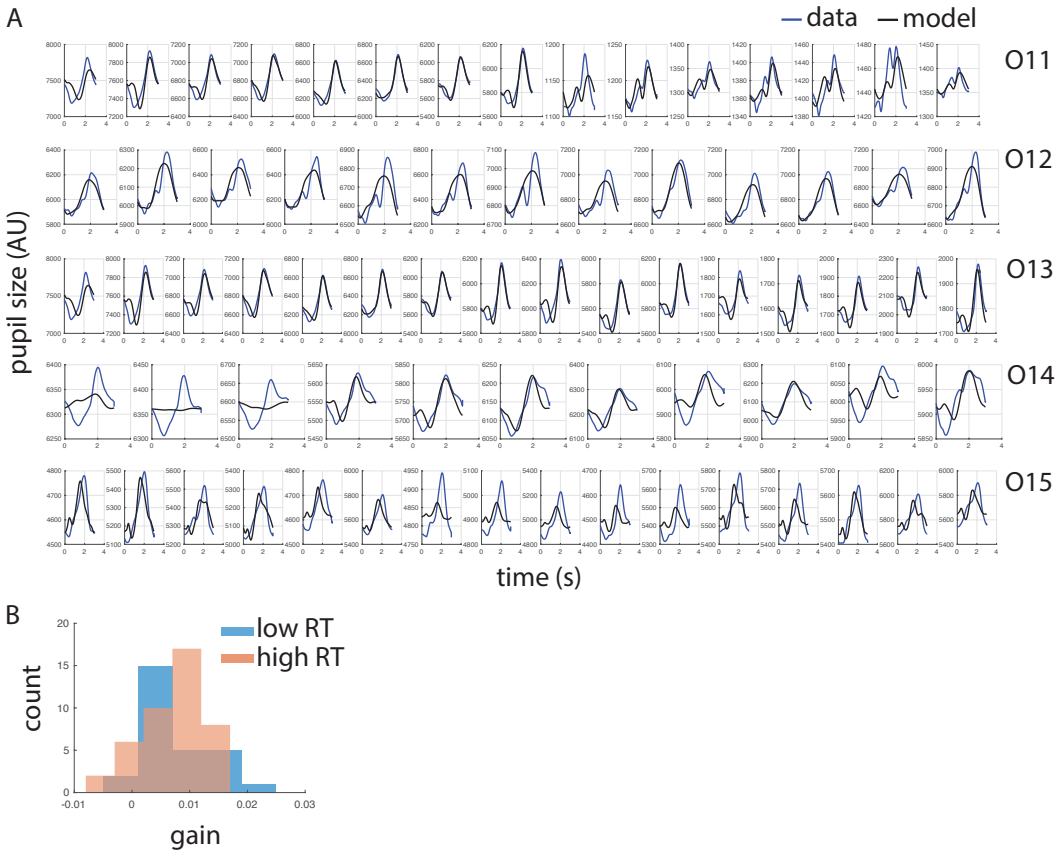


**Fig. S15.** Data and model fits from a version of the hard task with no staircase. The experimental protocol was identical to the one used in our main experiment (hard trials only), with two changes. (1) The tilt of the grating was fixed at each observer's pre-obtained 75% correct threshold. (2) Trials were 3 s in duration instead of 4 s. (**A**) Data and model fits for 5 new observers. Blue, task-evoked pupil response (N = 75 trials). Black, model fit. (**B**) Gain for high and low RT trials (median split of RT distribution), across all participants. This figure demonstrates that the model can provide good fits to data from five new participants (i.e., the model generalizes to new participants). And that it fits pupil data from a hard task in which the stimulus orientation is not staircased (i.e., when there is no uncertainty in the stimulus orientation).

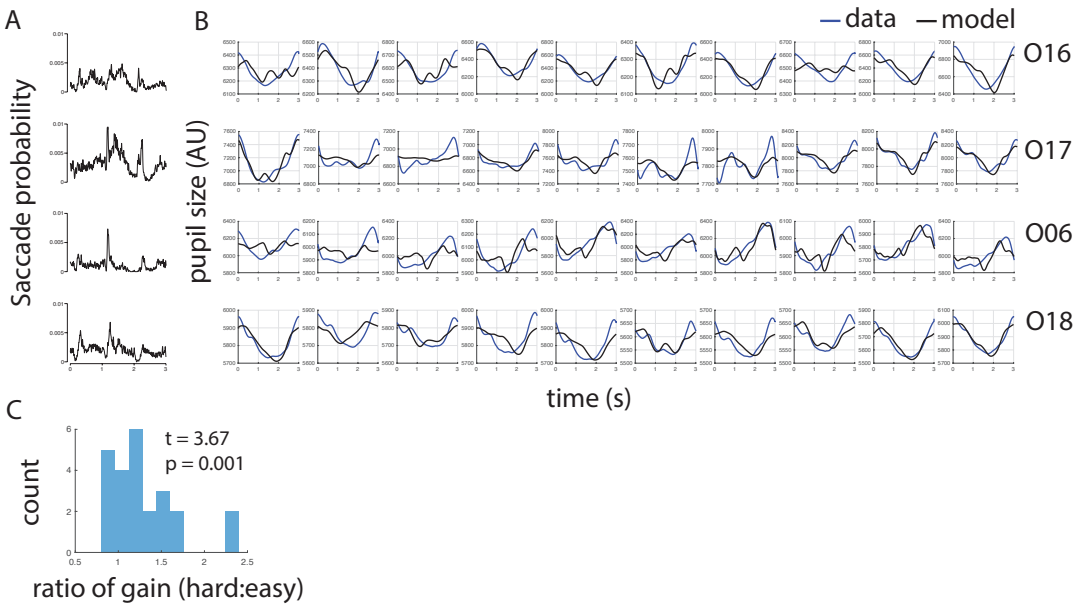


**Fig. S16.** Data and model fits from task with an annulus stimulus and task difficulty randomized across trials. The stimulus was a grating windowed by an annulus (outer diameter, 1.5 deg) around a central fixation dot. The tilt of the grating on each trial was drawn from a Gaussian distribution, whose standard deviation was determined by a thresholding procedure for each observer separately to achieve approximately 75% correct on average. The task structure consisted of a 900 ms ISI, a 200 ms stimulus presentation, and a 900 ms response window. (**A**) Saccade probability for each observer (N = 800 trials). Note that the average saccade rate over time is similar to that from the experiments presented in our paper, when the stimulus was peripheral, presented on the diagonal. (**B**) Task-evoked pupil response and model fit. Every subplot is for a different run of 160 trials. Same format as Fig. R5B. (**C**) Ratio of gain for hard vs. easy trials, where "hard" and "easy" were operationalized as low and high absolute stimulus tilts, using a median split of the distribution of tilts across trials. t and p statistics are from a paired t-test between 1 and the the distribution of gain ratios. We expect ratios greater than one under the hypothesis that task difficulty modulates arousal. This figure demonstrates that the model fits pupil data well from a task with a more complex trial structure (with a pre-cue), when task difficulty was randomly interleaved across trials rather than blocked, when the stimulus was presented in the fovea rather than the periphery, and for new observers. And it shows that gain was modulated by task difficulty to a similar degree as in our original experiment, but in the absence of a top-down cue indicating to observers the difficulty of the task.

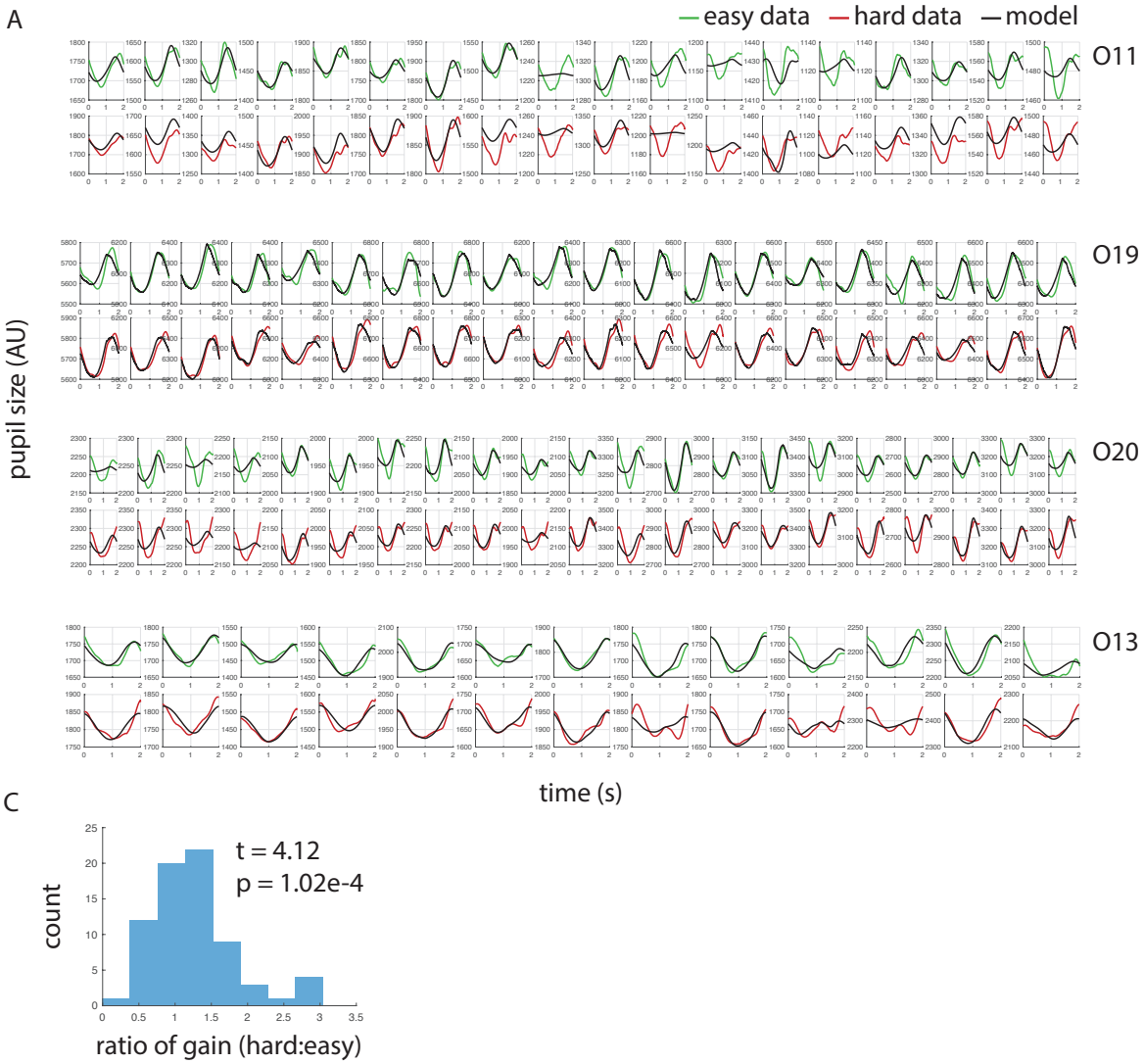


**Fig. S17.** Data and model fits for a task in which difficulty alternated between easy and hard every 5 trials. The stimulus and task was identical to the one described in the paper, except that task difficulty (and fixation color, green or red) alternated every 5 trials. Staircase values were carried over across hard blocks of trials. There were 160 trials per run, 2 s per trial. (**A**) Data and model fits for four new observers. Same format as Fig. R5B except that data from easy trials is in green and data from hard trials is in red. Each subplot is one run, with data aggregated over all easy or hard blocks within that run. (**B**) Ratio of gain for hard vs. easy blocks within each run. The t and p values are from a paired t-test vs. 1. We expect ratios greater than 1 under the hypothesis that task difficulty modulates arousal. This figure demonstrates that the model fits well when task difficulty is changed every 5 trials and for new observers.