Pupillary dynamics of optimal effort

Ceyda Sayalı[1], Emma Heling[2,3] and Roshan Cools[2,3]

[1]The Johns Hopkins University School of Medicine, Baltimore, MD
[2]Radboud University Medical Centre, Nijmegen, the Netherlands
[3]Donders Institute for Brain, Cognition and Behavior, Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands

Address correspondence to:
Ceyda Sayalı
5510 Nathan Shock Drive
Baltimore, MD 21224
tel 410-550-0100
email zsayali1@jh.edu

Running title: Optimal effort

ABSTRACT

While a substantial body of work has shown that cognitive effort is aversive and costly, a separate line of research on intrinsic motivation suggests that people spontaneously seek challenging tasks. According to one prominent account of intrinsic motivation, the Learning Progress Motivation theory, the preference for difficult tasks reflects the dynamic range that these tasks yield for minimization of performance accuracy prediction errors (Oudeyer, Kaplan & Hafner, 2007). Here we test this hypothesis, by asking whether greater engagement with intermediately difficult tasks, indexed by subjective ratings and objective pupil measurements, is a function of trial-wise changes in performance prediction error. In a novel paradigm, we determined each individual's capacity for task performance and used difficulty levels that are too low, intermediately challenging or high for that individual. We demonstrated that intermediately challenging tasks resulted in greater liking and engagement scores compared with easy tasks. Task-evoked and baseline pupil size tracked objective task difficulty, where challenging tasks were associated with smaller baseline and greater phasic pupil responses than easy tasks. Most importantly, pupil responses were predicted by trial-to-trial changes in expected accuracy, performance prediction error magnitude and changes in prediction errors (learning progress), whereas smaller baseline pupil responses also predicted greater subjective engagement scores. Together, these results suggest that what is underlying the link between task engagement and intermediate tasks might be the dynamic range that these tasks yield for minimization of performance accuracy prediction errors.

Keywords: Cognitive effort, pupil, expected accuracy, prediction error, learning progress

Introduction

According to the myth, Sisyphus was punished by the Gods to roll a boulder up a hill for eternity. It was the effort of his task that was considered a punishment for him. Although it is intuitive to assume that effort is aversive, people voluntarily engage throughout their lifespan in effortful tasks that allow them to acquire challenging hobbies, master expertise and be successful in adult life. Take video games for an example. Anyone who has played video games would agree that once they master a game level, they like to move onto a harder one. Most gamers would not voluntarily choose to play an easier level in order to avoid the effort. In fact, previous research (Baranes, Oudeyer & Gottlieb, 2014) has shown that when people are given the option, they gradually increase the difficulty of a video game across time. In these games, participants keep their task accuracy around 50% on average, even if they can choose to play the easier task level the entire time. An important question, then, concerns what underlies the preference for specific effortful tasks while others are deemed frustrating. In line with the theories on intrinsic motivation (Gottlieb & Oudeyer, 2018) as well as the recent proposal (Agrawal et al., 2021) suggesting that the utility of mental tasks increase as a function of the information they provide, we test the hypothesis that effortful tasks are preferred if they yield an opportunity for reduction of information uncertainty about performance success.

Influential theories of effort assume that cognitive effort holds an intrinsic cost (Shenhav, Botvinick & Cohen, 2013). As such, cognitive effort discounting and selection paradigms (Westbrook, Kester & Braver, 2013) show that participants forego reward to avoid the performance of challenging tasks. However, a separate line of research on motivation suggests that humans can be intrinsically motivated for optimal challenge. For example, intrinsically motivating activities that are intermediately challenging given one's capacity are known to induce a state of 'flow' (Csikszentmihalyi, 1990). In these activities, people report disliking tasks that are too easy or too difficult, but greater engagement with and liking of those tasks for which accuracy is around 50%. Moreover, neural networks recruited during the experience of flow correspond to regions that are commonly associated with reward receipt (Ulrich, Keller & Grön, 2016).

One prominent account that explains the mechanism underlying intrinsic motivation is Learning Progress Motivation theory (Oudeyer, Kaplan & Hafner, 2007). This account suggests that the motivational value of a task comes from its potential for performance improvement. Thus performance improvement might be proposed to correspond to the degree to which actual performance accuracy differs from the expected level of performance accuracy, that is, a performance accuracy prediction error, as well as its derivative (i.e. its change), which corresponds to the degree to which this performance accuracy prediction error is reduced across successive trials (Oudeyer, Gottlieb & Lopes, 2016).

Imagine three different types of task a researcher might need to perform for work. Say, one is a task they know how to execute by heart, such as subject data entry. They know they will perform this job with almost perfect accuracy, and they will succeed in doing so. This means their performance prediction errors for this task will be close to zero and will not change over time. Next, consider another task that is impossible for them to perform accurately, such as troubleshooting an fMRI machine. They know they will not be able to perform this task successfully, because they have no engineering background. So, their prediction error again is close to zero and stays flat. However, now consider a third job: a novel data analysis method that is intermediately challenging given their prior experience and associated with uncertainty about performance accuracy. They will fail once or twice, which will generate a negative performance prediction error given their expectation that they might well be correct. Upon persistence, they might succeed later on, which will boost their expectations about their own performance success, slowly reducing the difference between what they think they can do and what they actually can do.

According to Learning Progress Motivation theory, these intermediately challenging tasks are exactly the tasks in which internally motivated agents must invest effort. In a simulated experiment, researchers showed that artificial agents spontaneously spend more time exploring tasks that provide an opportunity for minimizing prediction errors and avoid tasks that yield no prediction error change (Gottlieb & Oudeyer, 2018). Similarly, human infants have been shown to attend to intermediately predictable auditory and visual stimuli (Kidd, Piantadosi & Aslin, 2014; Kidd, Piantadosi & Aslin, 2012). Based on this account, Sisyphus' punishment might have been reduced to the extent he experienced room for improving his skills in moving that boulder up the hill. In the current study, we test the hypothesis that intermediately challenging tasks are perceived as more engaging, as indexed both by subjective report of engagement as well as by trial-by-trial objective indices of pupil dilation, which has been well established to be associated with task engagement (Aston-Jones & Cohen, 2005). Furthermore, based on the Learning Progress Motivation hypothesis, we anticipate that the trial-by-trial index of pupil dilation is predicted by performance accuracy prediction error, as well as by its derivative, that is the change in performance accuracy prediction error.

To this end, we leverage a paradigm that is commonly used to induce a state of flow and engagement rather than avoidance of cognitive effort. After determining each individual's maximum cognitive capacity, participants performed 4 blocks of effortful tasks in which they scored 25, 50, 75 and 100% correct. After performance of this second effort exposure phase, participants completed self-report questions about their subjective liking, perceived ability and engagement during these different task blocks. We predicted that intermediately challenging tasks would be perceived as more engaging than easy or difficult tasks, yielding an inverted-U relationship between task engagement and task accuracy. We also predicted that subjective engagement scores would be predicted by both the average size of the performance accuracy prediction error as well as the average derivative (or change) of the performance accuracy prediction error across participants, with greater reductions in prediction error being associated with greater subjective engagement.

In addition, we acquired objective measures of engagement: the pupillary response during the task anticipation epoch of trials in the effort exposure phase. The pupil response has often been argued to reflect activity of the locus coeruleus (LC) (Murphy et al., 2014; Gilzenrat et al., 2010), the origin of noradrenaline, that is the neuromodulator most commonly implicated in task engagement (Aston-Jones & Cohen, 2005). Generally, an intermediate level of LC activity is considered optimal for task engagement, with both insufficient (boredom) and excessive arousal (stress) leading to impaired performance and reduced task engagement (Yerkes-Dodson curve (Yerkes & Dodson, 1908)). In fact, LC activity, and with it, the pupil response, has been shown to exhibit two modes of function: phasic and tonic. Phasic firing typically occurs in response to task-relevant events during epochs of high performance, and it is commonly characterized by large task-evoked dilations against a background of small baseline pupil size. This phasic mode is often contrasted with a tonic mode of pupil activity, which is associated with elevated baseline firing rate, absence of phasic responses, and degraded task performance (Aston-Jones et al., 1994; Aston-Jones & Cohen, 2005). By analogy, baseline pupil size was shown to be the highest and task-evoked dilations the lowest when participants decided to disengage from an effortful task (Gilzenrat et al., 2010) and this pattern reversed when participants reengaged with the task. Critically, in this prior work, decisions to engage were also always accompanied by higher accuracy, because participants were instructed to maximize accuracy-dependent reward. Therefore, participants chose to disengage from difficult tasks when their accuracy decreased too much.

Unlike this previous work on pupil dilation, the current task is predicted to dissociate task accuracy and reported task engagement, with participants putatively reporting greater subjective task engagement for tasks on which they perform more poorly than for tasks on which they perform perfectly. Therefore, our paradigm provides a novel tool not only for testing the Learning Progress Motivation hypothesis of motivation (Oudeyer, Kaplan & Lopes, 2007), but also for investigating more directly the putative selective link between pupil dilation and task engagement. Indeed, previously observed effects of task engagement on pupil response were often confounded by task difficulty (Beatty, 1982; see van der Wel & van Steenbergen, 2018 for a review). The present study addresses this confound.

Thus, two straightforward, dissociable predictions are tested: First, during the cue period, the phasic pupil mode will track participants' reported engagement, thus exhibiting a quadratic trend across task difficulty. Second, on a trial-by-trial basis, phasic pupil size will track the magnitude of the previous trial prediction error as well as previous prediction error change.

2. Methods

2.1. Participants

Forty English-speaking participants were recruited from the Radboud University participant pool (SONA Systems). All had normal or corrected-to-normal vision. They received a monetary reward (€10.00) for their participation and provided written informed consent prior to the experiment. Each participant was tested individually in a laboratory session lasting approximately 75 minutes. Participants were removed from the analyses if they had not completed the study (N = 2), or if no reliable pupil calibration accuracy could be obtained during the calibration phases (N = 2). The final sample size of 36 (ages 19-64; M = 24; *SD* = 9.6; 20 women see Supplementary Methods 1 for the age distribution and Supplementary Result 3 for analyses of a more homogeneous age-group) allowed us to detect an effect size of Cohen's d ≥ 0.05 with 80% power and alpha of 0.05 (Cohen's 1992).

2.2. Stimuli and Data Acquisition

During the computerized task portion of the experiment, participants were seated in a height-adjustable chair in front of a 23-inch monitor set to a resolution of 1920-1080 pixels, in a constant dimly lit room. Participants were instructed to keep their heads still and stabilized, rested in a chinrest positioned 50 centimetres away from the screen. All stimuli were delivered and controlled via the software Matlab (version 2016b) using Psychtoolbox library. Pupil responses were recorded using Eyelink 1000 eye tracker. Stimuli consisted of arithmetic summations of diverse difficulty levels, manipulated by summation length. Experimental scripts can be found on the author's personal Github page (https://github.com/zceydas/OptimalEffort_Pupillometry).

2.3. Procedure

2.3.1.The task

The participants started with a computerized task, which consisted of solving arithmetic summations varying in difficulty (e.g., 17 + 2) There were two different phases: a capacity phase to determine the four Task Difficulty levels, and a performance phase where those conditions are performed. Both will be explained in detail in the following sections. In both phases, participants had to solve summations by giving a free response within a time period of 18 seconds. They were instructed to answer as many trials correctly as possible, and to answer as soon as they knew the correct answer. The answer had to be entered digit by digit using numeric keys on the keyboard. Mistakes could be corrected using a “Backspace” button. Their input was immediately displayed on the screen. All summations were presented in a row, with accurate feedback (*correct*, *incorrect* or *too late*) provided immediately afterwards.

2.3.2.Capacity phase

The capacity phase served to determine the participants’ level of skill. Each difficulty level consisted of 5 summation trials. Starting at the easiest possible level, the difficulty level of the summations was continuously increased by one level. Specifically, a level-up adjustment occurred when the participant correctly answered at least one question of the same difficulty level. This procedure was continued until the participant scored 0% correct at all 5 summations of a certain difficulty level. When the capacity phase was finished, a sigmoid function was fitted to the resulting datapoints in order to estimate four Task Difficulty levels based on participant’s own accuracy. The ‘Easy’ condition was kept the same across all individuals and yielded the following formula: X + X. All participants scored 100% correctly on this task condition. The subsequent task levels yielded lower accuracy: At ‘Intermediate1’ condition accuracy was 75%, at ‘Intermediate2’ condition accuracy was 50%, at ‘Difficult’ level accuracy was 25%. These individually determined task conditions were used as the upcoming four Task Difficulty levels in the following performance phase.

2.3.3.Performance phase

In the performance phase, the participant completed two blocks per Task Difficulty in a pseudo-randomized order (Figure 1). Specifically, each Task Difficulty block was randomized in the first half of the performance phase (e.g. “C-A-D-B”), and the same order was repeated for the second half (“C-A-D-B”). Each block consists of eight trials, so that the entire performance phase comprised sixty four trials in total. Prior to each block, participants underwent a standard nine-point calibration procedure of gazing shortly at nine markers displayed on the screen sequentially, to ensure pupil capture calibration accuracy was kept similar across task blocks. At the beginning of each trial, a letter cue (A, P, Q, Y) was presented to indicate the upcoming task condition. These cue-task condition pairings were counterbalanced across participants in such way that each letter had a similar likelihood of being paired with a task condition across participants. This cue period lasted three seconds, to capture the entire response profile of the pupil. The cue was followed by a fixation cross of one second. Following the fixation cross, the participant solved a summation trial, the difficulty of which depended on the current Task Difficulty block. All cues, fixation crosses and summations were presented in red. Feedback consisted of a blue check mark for correct, an orangish cross for incorrect and an orange clock for too slow. All stimuli were presented against a green background. Stimuli and background colours were selected to keep constant the screen luminance (0.3*R + 0.59*G + 0.11*B = ~110 cd/m^2) as well as the contrast of the stimuli against the background during the whole task, to rule out any influence of screen luminance on pupil responses. Lastly, a final fixation cross appeared on the screen. Depending on response time (RT) in the summation period, the duration of this last fixation cross was adjusted, so that the total trial duration was always eighteen seconds. After each block, the participants provided a subjective flow rating, see below.

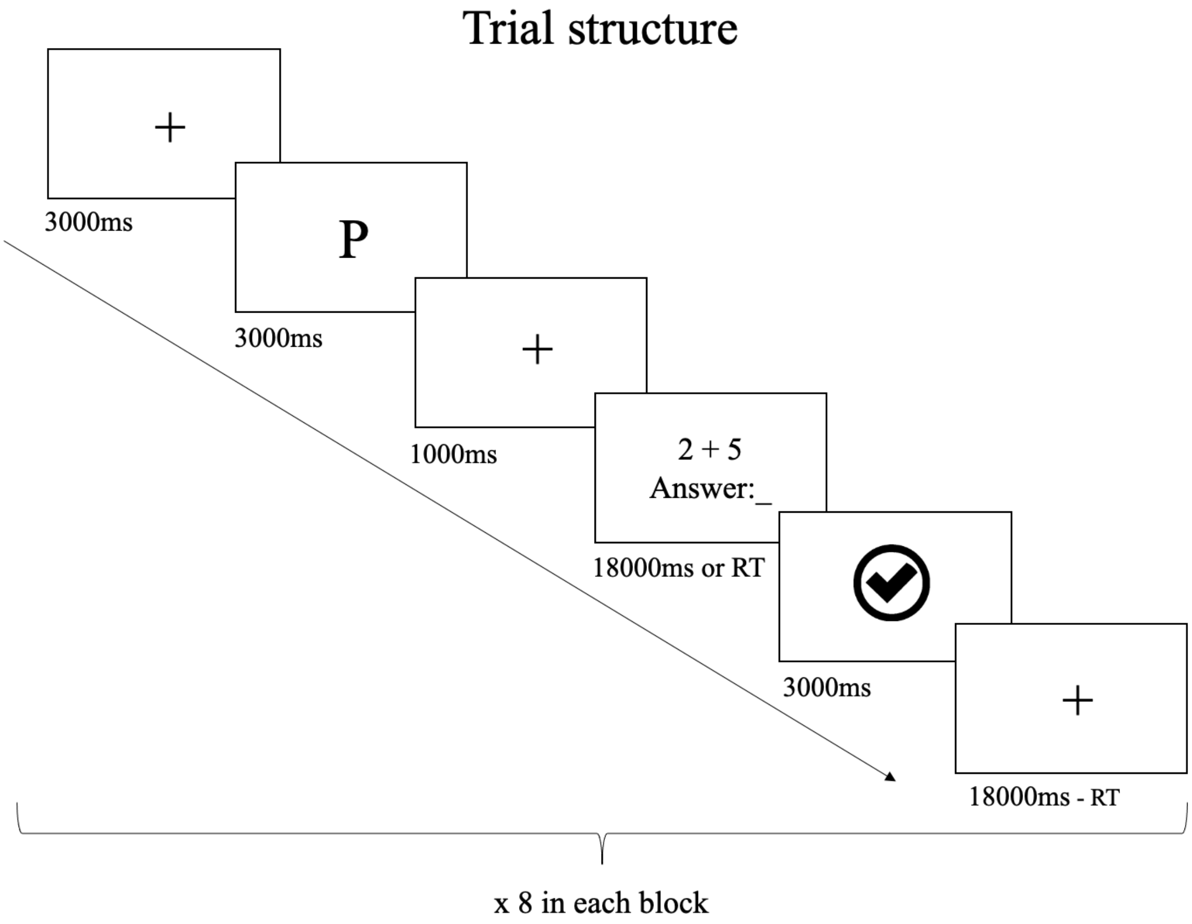


Figure 1. Illustration of one trial structure during the Performance Phase. The fixation cross preceding the cue presentation was 3 seconds. The cue was presented for 3 seconds and signalled that the upcoming task performance in the Performance Phase. In the Capacity Phase, there was no cue. After the presentation of a fixation cross for 1 second, participants solved the arithmetic summation question. The deadline for this epoch was 18 seconds. Upon response, participants received accurate feedback for 3 seconds. The final fixation cross duration was variable depending on RT only in the Performance Phase. During the Capacity phase, the final fixation duration was randomized with a mean of 2 seconds.

2.3.4.Subjective flow questionnaire

Subjective flow (flow index) was indexed by nine visual analogue ratings. The participant could range their response by moving the mouse on a horizontal line (10 cm in length) that had no anchors, except for the middle and endpoints. The answers were rated on a scale from 0 to 1. For 8 items, which measured the control component of flow, the endpoints were labelled *agree* and *disagree.* According to flow literature (Csikszentmihalyi, 1990; Ulrich et al., 2014; Ulrich, Keller & Grön, 2016), these items were used to monitor involvement, enjoyment, perceived fit between skills and task demands, and feeling of control with respect to each difficulty level (Table 1). A 9th statement assessed participants' subjective sense of time (Keller & Bless, 2008; Ulrich et al, 2014).

Q1. I would love to solve math questions of that kind
Q2. I was strongly involved in the task
Q3. I was thrilled
Q4. The task was boring
Q5. I had the necessary skills to solve the calculations successfully
Q6. Task demands were well matched to my ability
Q7. During the task all thoughts on task-irrelevant issues that I am personally concerned with were extinguished
Q8. During the task my consciousness was completely focused on solving the math calculations
Q9. The time passed really quickly

Table 1. Flow questionnaire items.

2.4.Statistical analyses

All statistical analyses were conducted in Rstudio (Version 1.3.1093). The analyses included analysis of variance (ANOVA) and Bayesian modelling. For the ANOVA, we used the R package *ez*. Bayesian models were created in Stan and assessed with *brms* package (Bürkner, 2017). For ANOVA, alpha level of 0.05 was used for all analyses. For Bayesian models, credible intervals with 95% probability were computed and parameters that did not include 0 within their credible intervals were considered significant in predicting the outcome variable.

2.4.1.Behavioural data analysis

2.4.1.1.Response time and accuracy rate

To validate the paradigm, mean accuracy rates (percentage correct trials) and mean response times for correct trials (seconds) from the performance phase were submitted to a one-way repeated measures ANOVA using Task Difficulty (easy, intermediate1, intermediate2, difficult) as a repeated measure. To assess simple effects of Task Difficulty, pairwise t-test analyses were executed with a Bonferroni correction. If the assumption of sphericity was not met, a Greenhouse-Geisser correction was applied (Field, 2009).

2.4.1.2.Flow measurement analysis

To address whether participants' flow ratings differ with Task Difficulty, we analysed the total flow score as well as four component scores (Supplementary Results 2): 'ability' (item 5 and 6), 'involvement (item, 2, 7 and 8), 'liking' (item 1, 3 and 4) and 'time' (item 9). Individual ratings were averaged across all trials from each task condition, for total flow as well as for each of the four component scores. Total flow and component scores were submitted to a repeated measures ANOVA with Task Difficulty level as a within-subject factor. Simple effects were investigated using Bonferroni corrected paired t-tests. In case of non-sphericity, a Greenhouse-Geisser correct was applied.

2.4.1.3.Model parameter analysis

To confirm that expected accuracy, performance accuracy prediction error (PE) and performance accuracy prediction error change (ΔPE) vary with Task Difficulty, we computed the following latent parameters: expected accuracy was computed as the cumulative sum of the probability of a correct answer on each trial (Nagase et al., 2018; Sayali & Badre, 2021), and updated across the two sessions for each Task Difficulty separately. PE was computed by subtracting expected accuracy from the corresponding accuracy value (1 for correct, 0 for incorrect; Figure 2). Finally, ΔPE reflects the difference between PE on two successive trials and indexes performance progress. These parameters were averaged across all trials per Task Difficulty level for every participant. For these validation analyses, averages of these parameters were submitted to a repeated measures ANOVA with Task Difficulty as a within-subject factor. Significant effects were followed by pairwise Bonferroni corrected t-tests. In case of violation of the sphericity assumption, a Greenhouse-Geisser correction was applied (Field, 2009).

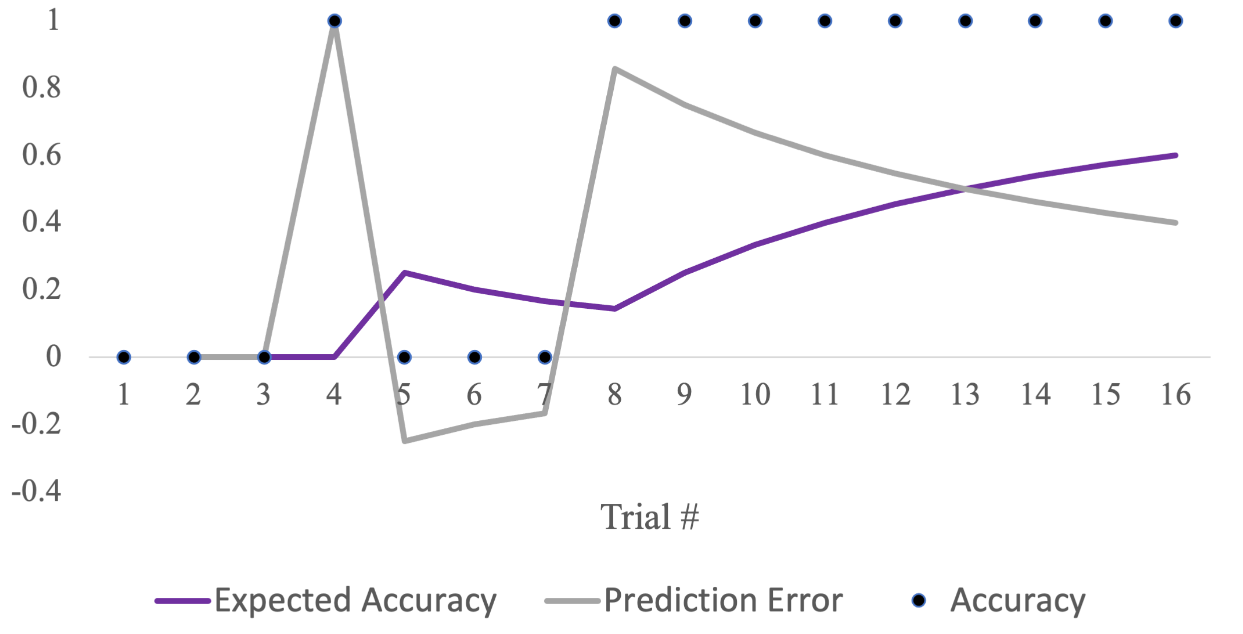


Figure 2. Exemplary progress of expected accuracy and prediction error parameters in relation to trial-by-trial changes in accuracy across all 16 trials (across two blocks) for the Difficult Task of one participant.

2.4.2. Pupillary data analysis

As in previous literature, pupil size, reported as pupil diameter, was registered during fixation, cue and feedback periods with a sampling rate of 500Hz. The obtained raw pupillometry data were exported and pre-processed in Matlab before calculating a trial-by-trial baseline and task-evoked pupillary response (TEPR) during letter cue presentation. First, across all task epochs (1st fixation, cue presentation, 2nd fixation and performance feedback epochs) we excluded 12.1% of the trials from further analysis based on our exclusion criterion of more than 40% NaNs per trial (*M* = 7.72%, *SD* = 5.96%). These reflected blinking. In the remainder of the data, missing data and eye blinks were detected (*M* = 5.8% of overall data, *SD* = 4.48%), removed and smoothed by convolution with a 11 ms hanning-window. The smoothed pupil recordings were corrected using cubic spline interpolation. After interpolation, the residual pupil time series were bandpass filtered using a 0.02-4 Hz third-order Butterworth filter, to decrease noise and remove slow drifts (Knapen et al. 2016). Next, we checked for effects of Task Difficulty on gaze drift in x- and y-direction with a linear mixed effects model. This gaze drift control analysis confirmed that the frequency of the saccades in both directions did not change as a function of Task Difficulty (GazeDrift_x: (F(3,34.33) = 0.89, p = .35) ; GazeDrift_Y: (F(1,34.32) = 0.99, p=.32)). Therefore, the effect of saccades from pupil responses were not removed. Subsequently, the time series were normalized within each block by z-scoring, in order to make comparisons between Task Difficulty blocks and to correct for individual differences in pupil diameter (de Gee, Knapen & Donner, 2014; Nassar et al., 2012, Urai, Braun & Donner, 2017). Time courses of pupil size changes before and after cue presentation are presented in Figure 3 to demonstrate the typical pupil response without baseline correction. Trial-by-trial baseline pupil diameter was calculated as the average unfiltered pupil diameter during the 200 ms period before cue onset. To control for variability in overall pupil size due to non-task related processes across trials within the same block, a baseline correction was applied to the standardized pupil units on a trial-by-trial basis by subtracting the preceding baseline pupil diameter (Eckstein, Starr & Bunge, 2019; Hershman & Henik, 2019) from the letter cue period. The final trial-by-trial TEPR was calculated as the maximum pupil diameter observed between a period of 1000 ms and 3000 ms after cue onset (Gilzenrat et al., 2010).

The pre-processed data were transported to Rstudio.

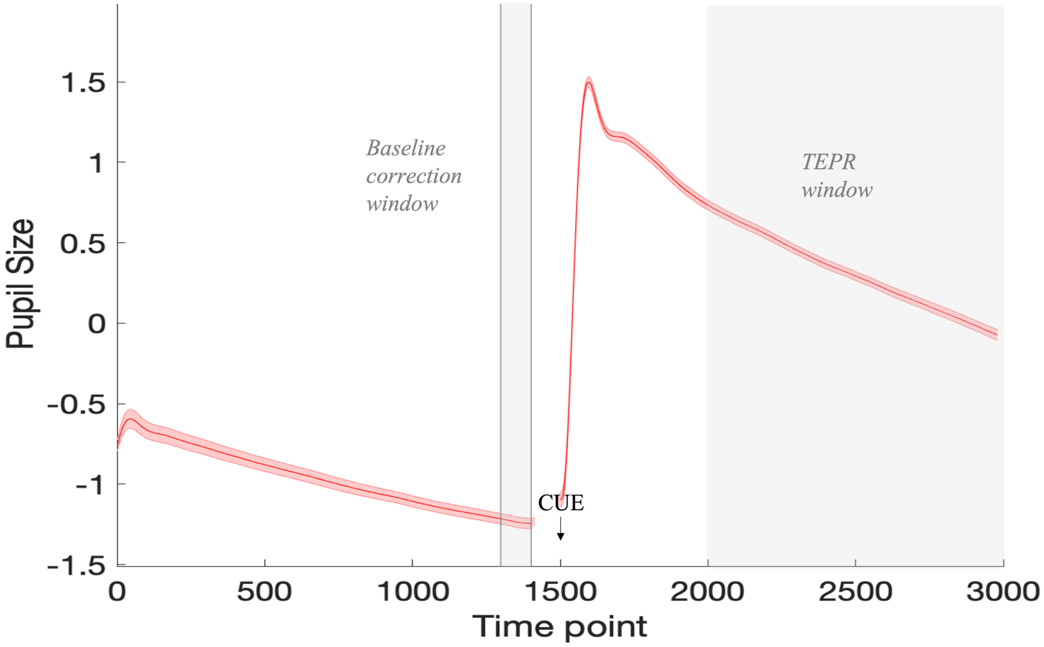


Figure 3 Average time course of (baseline uncorrected) pupil size across all task conditions from one and half seconds before and after cue onset.

The trial-by-trial cue and baseline data were analysed using a Bayesian mixed effects regression model (brms on R). The model predictors included the following covariates of interest: Task Difficulty and the latent model parameters PE, ΔPE and expected accuracy; as well as covariates of no-interest: trial number and task order, where trial number refers to the trial number within a task block (1 through 8) and task order refers to the order in which a task block is presented to the participant among other tasks in the experiment (1 through 8).

2.4.3.Pupillary – behaviour analysis

All parameters were estimated using Monte Carlo (MHC) with 5 chains of 2000 samples each. Significant predictors are determined by looking at the posterior distribution; a distribution not containing zero indicates a significant predictor. Participants' age, task order and trial number were included as nuisance variables in all relevant mixed effects models.

2.4.3.1.Effect of task related factors on pupil size

To probe the effect of task-related factors on pupil size, we separately examined the influence of expected accuracy, PE and ΔPE on Baseline and TEPR on a trial-by-trial basis. We used a Bayesian model comparison analysis of the non-averaged data, where subject-level parameters are drawn from group-level distributions. The models included the baseline and TEPR pupil size as dependent variables, with fixed effects and random slopes for expected accuracy, PE and ΔPE, Task Difficulty as well as nuisance variables trial number, task order and subject age, where subject number was entered as random effect (final model formula notation = (*PupilSize* ~ PE + ΔPE + ExpectedAccuracy + TaskDifficulty + TrialNo + TaskOrder + Age + (1 + PE + ΔPE + ExpectedAccuracy + TaskDifficulty + TrialNo + TaskOrder + Age || SubjectNumber, data=d, REML=F)). In addition, the model predicting TEPR also included the baseline pupil size in order to control for the effects of baseline correction in calculating TEPR.

Prior to each model fitting, predictor multicollinearity has been assessed by computing variance inflation factor (VIF), which measures the inflation of a regression coefficient due to collinearity between predictors (Bruce & Bruce, 2017; James et al., 2014). For example, a VIF above 5 is considered problematic and predictors that yield problematic VIF scores are considered redundant. For the model predicting baseline pupil size, variance inflation factor (VIF) index for all predictors was low ($VIF_{PE} = 2.67$, $VIF_{\Delta PE} = 2.79$, $VIF_{ExpectedAccuracy} = 2.11$, $VIF_{TaskDifficulty} = 2.07$, $VIF_{TaskOrder} = 1.07$, $VIF_{TrialNo} = 1.05$, $VIF_{Age} = 1.01$), justifying the inclusion of these terms in the same model. For the model predicting TEPR, variance inflation factor (VIF) index for all predictors was also low ($VIF_{Baseline} = 1.11$, $VIF_{PE} = 1.08$, $VIF_{\Delta PE} = 1.06$, $VIF_{ExpectedAccuracy} = 2.12$, $VIF_{TaskDifficulty} = 2.08$, $VIF_{TaskOrder} = 1.07$, $VIF_{TrialNo} = 1.05$, $VIF_{Age} = 1.00$).

2.4.3.2.Motivational relevance of pupil size

To test the motivational relevance of pupil size, we asked whether flow scores can be predicted by TEPR and baseline pupil size. To this end, we adopted a Bayesian mixed effects regression model approach to assess the non-averaged data. The models included total flow index as the dependent variable, with fixed effects as well as random slopes for baseline pupil size, TEPR, Task Difficulty and model parameters, expected accuracy, PE and ΔPE as well as nuisance variables trial number, task order and subject age, where subject number was entered as random effect (formula notation = (Flow ~ TEPR + Baseline + PE + ΔPE + ExpectedAccuracy + TaskDifficulty + TaskOrder + Age + (1 + TEPR + Baseline + PE + ΔPE + ExpectedAccuracy +

TaskDifficulty + TaskOrder + Age || SubjectNumber, data=d, REML=F)). Variance inflation factor (VIF) index for all predictors was low (VIF_{TEPR} = 2.68, $VIF_{Baseline}$ = 2.76, VIF_{PE} = 3.02, $VIF_{\Delta PE}$ = 2.94, $VIF_{ExpectedAccuracy}$ = 1.02, $VIF_{TaskDifficulty}$ = 1.03, $VIF_{TaskOrder}$ = 1.00, VIF_{Age} = 1.00), justifying the inclusion of these terms in the same model.

2.4.3.3.Behavioural relevance of pupil size

To test the behavioural relevance of pupil size, we assessed whether task performance can be predicted by TEPR, again using a Bayesian mixed effects regression model of the non-averaged data. The two separate models included current trial accuracy and RT as dependent variables, with fixed effects as well as random slopes for baseline pupil size, TEPR, Task Difficulty and model parameters, PE, ΔPE, expected accuracy and nuisance variables trial number, task order and subject age, where subject number was entered as random effect (formula notation = (*TaskPerformance* ~ Baseline + TEPR + PE + ΔPE + ExpectedAccuracy + TaskDifficulty + TaskOrder + TrialNo + Age + (1 + Baseline + TEPR + PE + ΔPE + ExpectedAccuracy + TaskDifficulty + TaskOrder + TrialNo + Age || SubjectNumber, data=d, REML=F)).

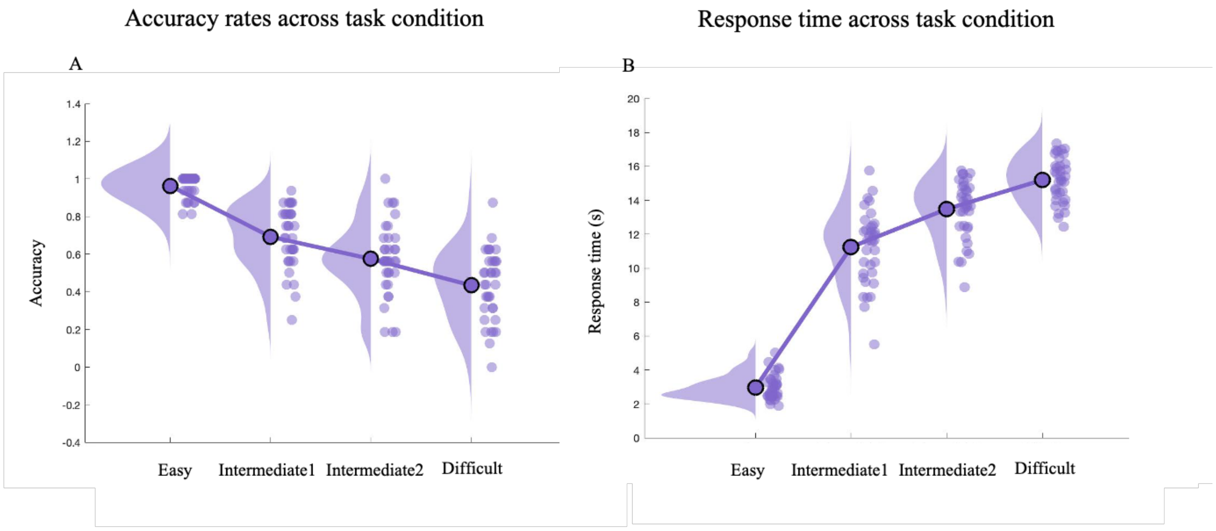
For the model predicting accuracy, the variance inflation factor (VIF) index for all predictors was low ($VIF_{Baseline}$ = 4.61, VIF_{TEPR} = 4.60, VIF_{PE} = 1.13, $VIF_{\Delta PE}$ = 1.11, $VIF_{ExpectedAccuracy}$ = 1.46, $VIF_{TaskDifficulty}$ = 1.44, $VIF_{TaskOrder}$ = 1.07, $VIF_{TrialNo}$ = 1.06, VIF_{Age} = 1.00), justifying the inclusion of these terms in the same model. For the model predicting correct RT, the variance inflation factor (VIF) indices for pupil predictors were high ($VIF_{Baseline}$ = 3.67, VIF_{TEPR} = 3.99, VIF_{PE} = 1.18, $VIF_{\Delta PE}$ = 1.09, $VIF_{ExpectedAccuracy}$ = 2.43, $VIF_{TaskDifficulty}$ = 2.51, $VIF_{TaskOrder}$ = 1.13, $VIF_{TrialNo}$ = 1.11, VIF_{Age} = 1.01), permitting the inclusion of all predictors in predicting correct RTs.

3. Results

3.1. Effect of Task Difficulty on task performance

Confirming our experimental manipulation, accuracy rate significantly decreased with Task Difficulty (*F*(3,105) = 106.639, *p* < .001) (Figure 3A). Accuracy at each effort condition was significantly different from the others (all *p*s<0.001).

Similarly, the time it took to respond correctly increased with Task Difficulty (*F*(2.267,79.335) = 755.848, *p* < .001) (Figure 3B). Average RT at each effort condition was significantly different from the others (all *p*s<0.001).



Accuracy rates across task condition

Response time across task condition

Figure 4. Probability density plots for A) accuracy rate, B) response time (RT) across task condition. The dots display each participant's mean accuracy rate and correct RT. Circled dots stand for group averages.

3.2. Effect of Task Difficulty on subjective flow scores

Although task performance declined monotonically with Task Difficulty, subjective engagement as indexed by the flow questionnaire increased with Task Difficulty. The flow index significantly differed across Task Difficulty (*F*(1.686,140) = 19.186, *p* < .001) (Figure 5). Both intermediate difficulty levels yielded greater flow scores compared with the Easy level (both *p*s<0.001) and Easy yielded greater flow scores compared with the Difficult level (p<0.001). No significant effects differences were found between Intermediate1 and Intermediate2 (*p* = 1.000), Intermediate1 and Difficult (*p* = 0.412, nor Intermediate2 and Difficult (*p* = 1.000), suggesting that the subjective experience of flow plateaued at the intermediate Task Difficulty levels. A closer look into the subcomponents of the flow inventory showed that task involvement and liking increased with Task Difficulty (see Supplementary Results 1), while perceived task ability decreased, indicating that task engagement, as indexed by the subjective experience of flow, dissociated from that of task ability.

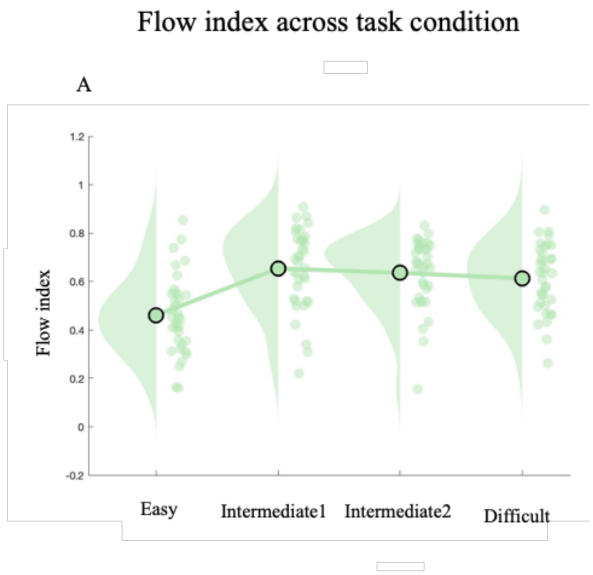


Figure 5. Raincloud plot for average flow scores across Task Difficulty.

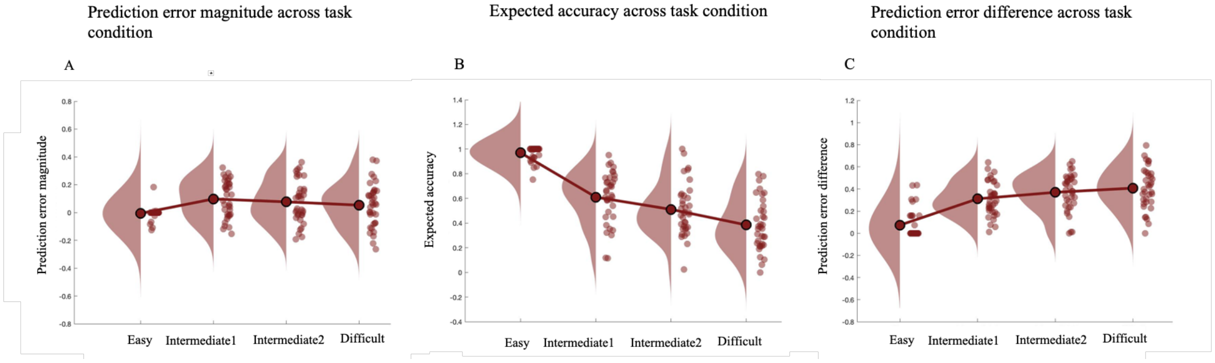
3.3. Effects of Task Difficulty on latent model parameters: expected accuracy and prediction error

Average expected accuracy (Figure 6) declined significantly with increasing Task Difficulty ($F(3,105) = 84.008$, $p < 0.001$) where Easy yielded significantly higher scores than Intermediate1 ($p < 0.001$), Intermediate2 ($p < 0.001$), and Difficult ($p < 0.001$). Moreover, average expected accuracy was also significantly higher in Intermediate1 vs Difficult ($p < 0.001$), and Intermediate2 vs Difficult ($p = 0.011$). No significant difference was observed between Intermediate1 vs Intermediate2 ($p = 0.180$).

Conversely, average prediction error (PE) magnitude differed significantly as a function of Task Difficulty ($F(2.517,88.095) = 4.862$, $p = 0.003$), but plateaued at the intermediate difficulty levels (Figure 6). Both Intermediate Task Difficulty levels yielded greater PE compared with Easy (Intermediate1: p<001; Intermediate2: p=0.024), but there was no difference between intermediate levels and Difficult or Easy vs Difficult (Easy vs. Difficult, $p = 0.169$; Intermediate1 vs

Intermediate2, p = 1.000); Intermediate1 vs Difficult, p = 0.915; Intermediate2 vs Difficult, p = 1.000).

Finally, ΔPE increased with increasing Task Difficulty (F(3,105) = 27.938, p < 0.001) (Figure 5C), also reaching a plateau at intermediate levels: pairwise contrasts were significant between Easy versus both Intermediate Task Difficulty (both, ps < 0.001), and versus Difficult (p < 0.001), but with no significant differences between Intermediate1 vs Intermediate2 (p = 0.790), Intermediate1 vs Difficult (p = 0.123), or Intermediate2 vs. Difficult (p = 1.000).


Prediction error magnitude across task condition

Expected accuracy across task condition

Prediction error difference across task condition

Figure 6. Raincloud plot for A) prediction error (PE), B) expected accuracy, and C) prediction error difference (ΔPE). The dots display each participant's mean PE, expected accuracy, and ΔPE.

3.4. Effects of task-related factors on pupil size

To assess which (trial-by-trial latent) processes/factors contribute to trial-by-trial changes in baseline and TEPR, we ran a mixed effects model with predictors of interest, Task Difficulty, expected accuracy, PE and ΔPE and the nuisance regressors, trial number, task order and subject age.

Baseline pupil size on the current trial *decreased* with PE (B = -0.56, CI [-0.65, -0.48]), ΔPE (B = -0.11, CI [-0.17, -0.05]), expected accuracy (B = -0.32, CI [-0.47, -0.17]) and Task Difficulty (B = -0.15, CI [-0.18, -0.11]), indicating that smaller baseline pupil size was associated with greater prediction error, learning progress, expected accuracy and task difficulty (Figure 7A, 8A).

On the other hand, pupil size during the cue period (TEPR) was largely predicted by the baseline pupil size. TEPR on the current trial *increased* with smaller baseline pupil size (B = -0.88, CI [-0.91, -0.85]), greater expected accuracy (B = 0.17, CI [0.09, 0.26]) and easier task blocks

(Task Difficulty: (B = -0.12, CI [-0.14, -0.10])). The effect of PE and ΔPE were not significant (PE (B = -0.02, CI [-0.06, 0.02]), ΔPE (B = -0.01, CI [-0.04, 0.02]) In other words, phasic pupil response showed an inverse relationship with the baseline pupil size and was greater when the expected accuracy of the upcoming trial was higher (Figure 7B, 8B).

These results indicate that baseline pupil size was smaller with increasing expected accuracy, being better than expected and showing greater performance improvements and smaller baseline pupil sizes predicted greater phasic pupil response, displaying an inverse relationship.

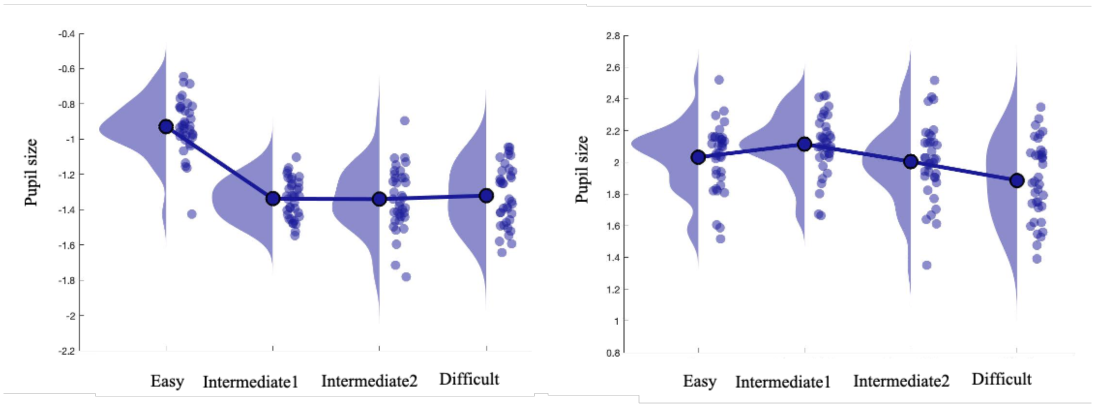


Figure 7. A) Average Baseline pupil size across Task Difficulty. B) Average TEPR across Task Difficulty.

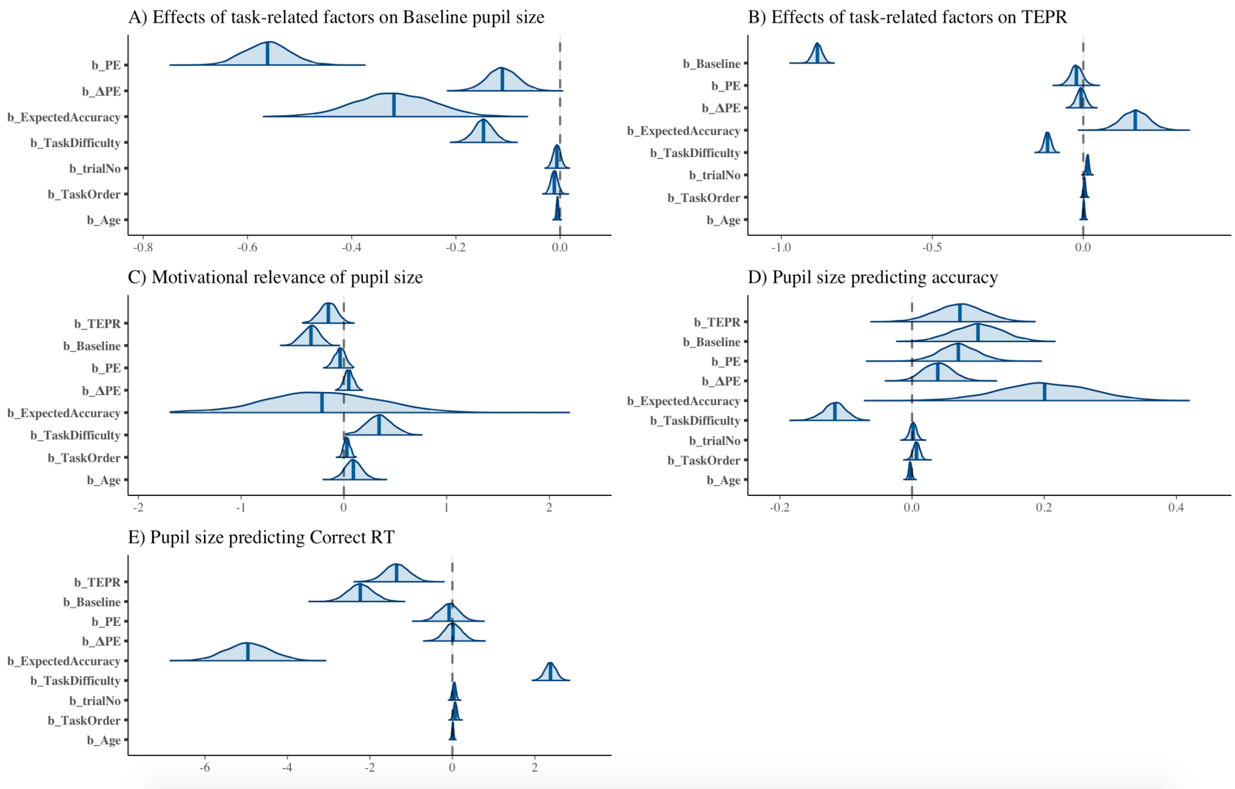


Figure 8. Densities of model parameter estimates. A) Model parameter densities for predicting Baseline pupil size. B) Model parameter densities for predicting TEPR. C) Model parameter densities for predicting flow scores. D) Model parameter densities for predicting accuracy. D) Model parameter densities for predicting Correct RTs.

3.5. Motivational relevance of pupil size

Next, we assessed the motivational relevance of pupil size by testing whether pupil size predicts subjective flow scores. Note that flow questionnaire responses were collected at the end of each task block and not on a trial-by-trial basis. Accordingly, the following model did not include the nuisance regressor *trial number.* The full model included the predictors of interest, *TEPR, baseline pupil size, PE,* ΔPE, expected accuracy*,* Task Difficulty *and* the nuisance regressor *task order and subject age*.

The model results (Figure 8C) showed that Baseline pupil size and Task Difficulty were significant predictors of subjective flow scores. Participants reported greater flow scores following trials with smaller baseline pupil size (B = -0.32, CI [-0.47, -0.16]) and greater Task Difficulty (B = 0.30, CI [0.08, 0.57]). The effect of other predictors were not significant (TEPR (B = -0.16, CI [-0.31, 0.00]), PE (B = -0.02, CI [-0.14, 0.07]), ΔPE (B = 0.06, CI [-0.03, 0.13]); expected accuracy

(B = -0.20, CI [-1.29, 0.84])) indicating that greater subjective engagement was associated with smaller baseline pupil size and greater task difficulty.

3.6. Behavioural relevance of pupil size

Finally, we assessed the behavioural relevance of pupil size by testing the influence of pupil size on upcoming task accuracy and correct trial RTs. This model included the predictors of interest *TEPR, baseline pupil size, PE,* ΔPE, *expected accuracy* and *Task Difficulty* and the nuisance regressors *task order, trial number and subject age*.

The model results predicting accuracy (Figure 8D) showed that the significant predictors of trial-by-trial accuracy were baseline pupil size (B = 0.10, CI [0.02, 0.17]), PE: B = 0.07, CI [0.01, 0.13], expected accuracy: B = 0.20, CI [0.07, 0.33]) and Task Difficulty (B = -0.12, CI [-0.15, -0.09]), while the effect of remaining predictors were not significant (TEPR: B = 0.07, CI [-0.00, 0.14]; ΔPE (B = 0.04, CI [-0.00, 0.08]). As expected, accuracy was higher on trials in easier task blocks and following trials with better performance than expected. Importantly, baseline pupil size showed a positive correlation with accuracy. As such, accuracy on the upcoming trial was higher when baseline pupil size was greater, while TEPR did not correlate with accuracy.

The model results predicting correct RTs (Figure 8E) showed that the significant predictors of trial-by-trial RT were TEPR (B = -1.36, CI [-1.94, -0.79]), baseline pupil size (B = -2.24, CI [-2.83, -1.66]), expected accuracy (B = -4.97, CI [-6.02, -3.85]) and Task Difficulty (B = 2.38, CI [2.11, 2.64]). The effect of remaining predictors were not significant (PE: B = -0.08, CI [-0.55, 0.39]; ΔPE (B = 0.02, CI [-0.39, 0.43]). As expected, RTs was longer on trials in more difficult task blocks and when the expected accuracy was lower. Moreover, both smaller baseline pupil size coupled with smaller TEPRs predicted longer RTs.

Discussion

While a substantial body of work has shown that cognitive effort is aversive and costly (Kool et al., 2010; Westbrook, Kester & Braver, 2013; Sayali & Badre, 2020), a separate line of research on intrinsic motivation suggests that people spontaneously seek challenging tasks. According to one prominent account of intrinsic motivation, the Learning Progress Motivation theory, the preference for difficult tasks reflects the dynamic range that these tasks yield for minimization of performance accuracy prediction errors (Kaplan & Oudeyer, 2007; Oudeyer, Gottlieb & Lopes, 2016). Here we test this hypothesis, by asking whether greater engagement with intermediately difficult tasks, indexed by subjective ratings and objective pupil measurements, is a function of trial-wise changes in performance prediction error.

Across four individually assigned difficulty levels, the current study dissociated subjective task engagement from task difficulty. As such, the current results show that subjective engagement and liking scores, as indexed by the flow questionnaire, increased with increasing task difficulty despite significant decreases in task performance and perceived task ability. Previous research (Ulrich et al., 2014; Ulrich, Keller & Grön, 2016; Katahira et al., 2018) on flow states has shown that these states could be induced empirically by dynamically adjusting the difficulty of the task to match participants' current performance levels. For example, in these flow conditions, if the participant correctly solved the last two trials, the difficulty of the current trial was increased by one level, resulting in an average of 50% accuracy across trials. This stood in contrast to the experience of easy and difficult task levels, in which the accuracy was deterministically above 90% or less than 5%, respectively. Thus, in contrast to the experience of overly easy and overly difficult tasks, where participants were certain that their performance would either almost always be correct or incorrect, during flow conditions at least two task parameters were different: 1) participants were not certain of the outcome of their performance ('performance uncertainty'); and 2) participants' performance on the current trial had a causal effect on the difficulty of the next trial. With the current study, we isolated the role of performance uncertainty on task engagement by assigning difficulty levels corresponding to 25, 50, 75 and 100% correct. Thus, critically, unlike the classic flow paradigm, our conditions differed only in terms of performance uncertainty, while keeping constant other factors of no interest.

In this updated design, we showed that intermediate effort levels yielded greater prediction errors as well as prediction error changes compared with the easiest difficulty level but not the most difficult level, leading to a plateau of prediction error changes as well as subjective engagement scores at the Intermediate1 difficulty level. This plateau effect stands in contrast to previous flow findings (Ulrich et al., 2014; Ulrich, Keller & Grön, 2016), which showed an inverted-U shape function of subjective task engagement across difficulty levels. We argue that this discrepancy might be due to differences in study designs. We found that learning progress (prediction error minimization) was the highest in the most difficult task condition, indicating that participants showed the greatest task improvement at the task level they were initially around %25 correct, increasing average accuracy ~40% accuracy across two blocks. The most difficult task was still within the capacity range of the participants, and hence, was not impossible for them to master. This is in contrast with Ulrich et al. (2014, 2016)'s 'overload' task manipulation in which the task level was way above the participant's own capacity and the average accuracy rate was around %5. This difference between paradigms made the most difficult task condition in our study another intermediately challenging task condition. Importantly, we found that subjective engagement also increased with increasing effort level, prediction error magnitude and prediction error minimization (learning progress). Thus, we have shown that intermediately challenging tasks which yielded the greatest range for prediction error changes received greater liking and engagement scores than the easiest level, leading to plateau of subjective pleasure across difficulty levels. These results suggest that what might be underlying the subjective pleasure during flow states might be the associated increased dynamic range for minimization of performance accuracy prediction errors.

This observation is generally consistent with previous literature indicating that perceived effort costs can be alleviated by subjective reward received during task performance (Inzlicht, Shenhav & Olivola, 2018). Furthermore, Devine and Otto (2021) have shown that receiving temporal information about progress reduced demand avoidance in a demand selection task in the absence of reward, indicating that information regarding performance improvements influenced cost-benefit decisions regarding cognitive effort. These results underscore the value of progress in mediating effort costs. Consistently, Geana et al. (2016) have manipulated the predictability of a task by asking the participants to predict numbers generated by a virtual machine and controlling the difference between the predicted number and the actual generated number (prediction errors). They have shown that people switched from a current task when that current task offered too much or too little change in prediction errors, indicating that an optimal amount of information gain, as tracked by prediction errors, is necessary for the motivation to stay on task. Indeed, it was recently proposed that easy tasks that are below participants' own capacity are perceived as boring, because they offer little to no information gain (Agrawal et al., 2021). This framework assumes that mental tasks occupy limited resources and engaging in such tasks incurs an opportunity cost of not engaging in other potentially valuable tasks (Kurzban et al., 2013). Thus, the utility of the current task signals whether to stay engaged with the current task or explore other tasks to maximize reward in a given environment. Further, the utility of a task is comprised of both its expected rewards and the value of the information it provides. Thus, an agent who mastered a task would be expected to switch to other tasks that provides opportunity for learning. This prediction concurs with the finding that (Baranes, Oudeyer & Gottlieb, 2014), in a video game setting, when people are given the option to either repeat the same difficulty level or to voluntarily increase difficulty, they gradually increase the difficulty of the game they play even when it means having poorer task performance. The current study corroborates these observations and firmly establishes the link between effortful task engagement and (the reduction of) performance prediction error.

Our findings are reminiscent of studies suggesting a key role for efficacy in the willingness to exert cognitive effort. Efficacy refers to how much one's efforts will impact the outcome of these efforts (Bandura, 1986). For example, if an outcome is driven mostly by factors outside of one's control, one's efficacy would be low. An updated version of the Expected Value of Control theory (EVC; Shenhav, Botvinick & Cohen, 2013) incorporates self-efficacy as a determinant of cost-benefit calculations regarding effort allocation, where higher efficacy increases the amount of effort exertion (Blain & Sharot, 2021; Frömer et al., 2021; Shenhav, Botvinick, & Cohen, 2013). Consistent with this framework, a recent study manipulated subjective efficacy by changing the contingency between actions and reward outcomes in a Stroop task design (Frömer et al., 2021). They demonstrated that if participants perceived their efforts as efficacious, they are more likely to exert control, as indexed by higher contingent negative variation (CNV) amplitude, an event-related potential (ERP) known to track proactive control allocation, and higher P3b, ERP known to signal incentive evaluation, during initial cue period. The current study goes beyond that prior work by decoupling task difficulty from outcome contingency: none of the task conditions received performance-contingent rewards. More specifically, our findings substantiate Learning Progress Motivation theory, according to which minimization of prediction errors registers as value. These results might thus inspire the updating of current resource allocation models of cognitive effort, such as EVC, with parameters capturing performance accuracy prediction error and/or prediction error change.

The current study results raise the possibility that increases in challenging task engagement are accompanied by prediction error-related changes in arousal, perhaps mediated by locus coeruleus (LC) activity and noradrenaline.. According to Adaptive Gain Theory of LC (Aston-Jones & Cohen, 2005), there are two modes of LC activity: tonic and phasic. Baseline and short-term burst-like activities of LC are typically inversely correlated. In a tonic mode, baseline LC activity is greater than momentary LC bursts. In this mode, the system is disengaged, and behavior is inattentive or distractible. On the other hand, the phasic mode is characterized by low baseline LC activity and greater LC bursts that are typically coupled with task-relevant outcomes or responses. Thus, in order to verify the subjective preference of challenging tasks, we related pupil size, as an objective marker of task engagement to subjective engagement scores.

Previous research has shown that phasic pupil response, which has been demonstrated to correlate with LC activity (Gilzenrat et al., 2010), tracks task utility where reward was coupled to task performance. The results were interpreted to reflect disengagement from tasks that were too difficult. In line with these observations, the current results show that in easier tasks where expected accuracy was higher were accompanied by smaller baseline pupil size and greater TEPRs also in the absence of performance-contingent rewards. These results indicate that expected performance mediates pupil responses in a way similar to expected reward and are consistent with previous research (Massar et al., 2016) which showed that an increase in pupil diameter was only present when rewards were contingent on good performance (high reward condition) but not when reward was provided at random (random reward condition), indicating the coupling between pupil size and performance-based reward might be mediated by expected task accuracy.

More critically, baseline pupil size were also predicted by performance prediction errors (PE) (i.e. being better than expected) as well as performance progress (ΔPE): Greater PE and greater performance progress were associated with smaller baseline pupil size. Previous evidence indicates that pupil size increases monotonically with greater task difficulty (Belayachi et al., 2015; Brouwer, et al., 2014; Irons, Jeon & Leber 2017; Klingner, Tversky & Hanrahan, 2011; Moresi et al., 2008, van der Wel & van Steenbergen, 2018). Our results demonstrate that task-evoked pupil size changes with engagement, in addition to task difficulty or accuracy changes. This conclusion is generally in line with recent evidence presented by da Silva Castanheira, LoParco, and Otto (2020), who demonstrated that task-evoked pupil size changes as a function of effort investment, even when task demands were kept constant. Their finding that pupil size was associated with better task performance, despite constant task demands, led them to propose that the pupil response serves as a reliable index of cognitive effort investment. Here we go beyond this by showing that this link exists even in the presence of accuracy decline, and can be accounted for, in part by changes in performance accuracy prediction error (changes).

Note that, contrary to baseline pupil size, cue-related pupil size did not relate to PE or ΔPE and only tracked expected accuracy of the upcoming trial. These results suggest that the prediction error related to the previous task performance no longer influences the pupil size once the new trial is signaled by the presentation of the cue. Importantly, the current study tested the relationship between task difficulty and phasic pupil size (TEPR) during a cue period and not during task performance. As such, cue related pupil size only tracked expected accuracy of the upcoming task, consistent with previous accounts (Kurniawan, Grueschow & Ruff, 2021) which dissociated the time of effort anticipation from that of effort exertion in an instrumental effort paradigm and showed that pupil size tracked the anticipation of the upcoming effort level of the task in a voluntary choice paradigm. Moreover, these increases in pupil dilation were stronger when the participants accepted to exert the effort option versus not, suggesting that pupil size might be associated with the energization required to perform an upcoming action. Current results, combined with the previous literature, suggests that cue related pupil size signals expectations about effort requirements rather than the exertion of effort itself.

The link between noradrenergic arousal and task engagement has been argued to be mediated by the anterior cingulate cortex (ACC). The ACC is a primary source of cortical input to the LC, and encodes uncertainty-related signals that correlate with pupil diameter, a putative proxy of LC activity. Intriguingly, Muller and colleagues (2019) have recently shown that changes in the perceived uncertainty in the internal model of environmental states might be linked to changes in noradrenergic functioning. Pupil diameter was larger during periods of perceived uncertainty and constricted as expectations became more reliable. ACC activity correlated with these trial-by-trial changes in pupil size. Furthermore, another study (Lavin, San Martín & Rosales Jubal, 2014) found that pupil size tracked task uncertainty and surprise independent of feedback magnitude in a reward learning task, suggesting that pupil size might be a marker of learning in uncertain environments. The current results corroborate these findings by showing that pupil size tracked performance prediction errors and their minimization independent of task difficulty, potentially boosting task engagement in more challenging tasks via noradrenergic arousal mechanisms. The trial-by-trial relationship between prediction error change and pupil size, provides direct evidence for the Learning Progress Motivation theory (Oudeyer, Gottlieb & Lopes, 2016), which states that motivation for challenging tasks is a function of the opportunity for improving performance and learning.

These findings also parallel evidence that the experience of flow is characterized by an intermediate level of arousal, as indexed by intermediate sympathovagal system activation (heart rate and breathing rate) and an intermediate level of ACC activity (de Sampario Barros et al., 2018; Ulrich et al., 2014; Ulrich, Keller & Grön, 2016). Moreover, direct stimulation of the vagus nerve, potentially via its role in activating LC and modulating noradrenaline release, was shown to be associated with reports of reduced flow experience (Colzato, Wolters & Peifer, 2018). However, no studies to date have directly tested the relationship between the LC-associated pupil response, performance predictions and optimally challenging effort. The current study firmly establishes such a relationship between pupil size and challenging task anticipation by showing that smaller baseline pupil size predicted greater subjective task engagement as assessed by the flow questionnaire at the end of each task block, while TEPR did not relate to flow scores. The lack of TEPR-flow relationship in the presence of baseline pupil-flow relationship might underscore the dissociation between baseline pupil size and TEPR in tracking motivational salience and task preparation, respectively (Chiew & Braver, 2013; Kostandyan et al., 2019).

While the mechanism underlying intermediate effort preference in our study points to the role of LC-based arousal, pupil size also has been associated with dopaminergic activity (Manohar & Husain, 2015; Muhammed et al., 2016). The dopaminergic system is mostly located in the midbrain ventral tegmental area (VTA) and substantia nigra (SN), and is traditionally involved in reward processing and value-based behaviour (Olds & Milner, 2020; Schultz, Dayan & Montague, 1997; Schultz, 2007). Moreover, DA is involved in tracking reward uncertainty, in a way to facilitate learning (Gershman & Uchida, 2019; O'Doherty et al., 2003). Future studies should disentangle the role of noradrenergic and dopaminergic systems in mediating the relationship between intermediate challenge and pupil size.

The interpretation of the results also are not without constraints. As described earlier, unlike in the flow induction paradigm (Ulrich et al., 2014; Ulrich, Keller & Grön, 2016), our design did not include a supra-threshold capacity difficulty level. Thus, in our task, all task conditions except the easiest difficulty level could be mastered by the participant. Therefore, the current design cannot answer what factors underlie the (dis)engagement during tasks that are above capacity limits.

Finally, although we measured subjective and pupillary engagement in the absence of external reward, in order to induce a state of flow, we provided performance feedback (Csikszentmihalyi, 1990). Hence, we were not able to test whether these findings would also hold in the absence of external feedback, which might be considered a form of external reward, underlie individual variability in feedback sensitivity and complicate our definition of internal rewards.

References

Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2021). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. Psychological Review.

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.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience, 28,* 403-450. Doi:10.1146/annurev.neuro.28.061604.135709.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology*, *4*(3), 359-373.

Baranes, A. F., Oudeyer, P. Y., & Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience*, *8*, 317.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Belayachi, S., Majerus, S., Gendolla, G., Salmon, E., Peters, F., & Van der Linden, M. (2015). Are the carrot and the stick the two sides of same coin? A neural examination of approach/avoidance motivation during cognitive performance. *Behavioural Brain Research*, *293*, 217-226.

Blain, B., & Sharot, T. (2021). Intrinsic reward: potential cognitive and neural mechanisms. *Current Opinion in Behavioral Sciences*, *39*, 113-118.

Brouwer, A. M., Hogervorst, M. A., Holewijn, M., & van Erp, J. B. (2014). Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *International Journal of Psychophysiology*, *93*(2), 242-252.

Bruce, P, & Bruce, A. (2017). *Practical Statistics for Data Scientists*. O'Reilly Media.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using. Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Chiew, K. S., & Braver, T. S. (2013). Temporal dynamics of motivation-cognitive control interactions revealed by high-resolution pupillometry. *Frontiers in psychology*, *4*, 15.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98-101.

Colzato, L. S., Wolters, G., & Peifer, C. (2018). Transcutaneous vagus nerve stimulation (tVNS) modulates flow experience. *Experimental Brain Research*, *236*(1), 253-257.

Csikszentmihalyi, M. (1990). The domain of creativity. In M.A. Runco & R. S. Albert (Eds.), *Theories of creativity* (pp. 190-212). Newbury Park, CA: Sage.

da Silva Castanheira, K., LoParco, S., & Otto, A. R. (2020). Task-evoked pupillary responses track effort exertion: evidence from task-switching. *Cognitive, Affective, & Behavioral Neuroscience*, 1-15.

de Sampaio Barros, M. F., Araújo-Moreira, F. M., Trevelin, L. C., & Radel, R. (2018). Flow experience and the mobilization of attentional resources. Cognitive, Affective, & Behavioral Neuroscience, 18(4), 810-823.

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-E625.

Devine, S., & Otto, A. R. (2021). Information about task progress modulates cognitive demand avoidance.

Eckstein, M. K., Starr, A., & Bunge, S. A. (2019). How the inference of hierarchical rules unfolds over time. *Cognition*, *185*, 151-162.

Field, A. (2009). Discovering Statistics Using SPSS, Third Edition. *Sage Publications Limited.*

Frömer, R., Lin, H., Wolf, C. D., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(1), 1-11.

Geana, A., Wilson, R., Daw, N. D., & Cohen, J. (2016). Boredom, Information-Seeking and Exploration. In CogSci.

Gershman, S. J., & Uchida, N. (2019). Believing in dopamine. *Nature Reviews Neuroscience*, *20*(11), 703-714.

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.

Gottlieb, J., & Oudeyer, P. Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758-770.

Hershman, R., & Henik, A. (2019). Dissociation between reaction time and pupil dilation in the Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1899.

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences*, *22*(4), 337-349.

Irons, J. L., Jeon, M., & Leber, A. B. (2017). Pre-stimulus pupil dilation and the preparatory control of attention. *PLOS one, 12*(12). Doi:10.1371/journal.pone.0188787.

Kaplan, F., & Oudeyer, P. Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience*, *1*, 17.

Katahira, K., Yamazaki, Y., Yamaoka, C., Ozaki, H., Nakagawa, S., & Nagata, N. (2018). EEG correlates of the flow state: A combination of increased frontal theta and moderate frontocentral alpha rhythm in the mental arithmetic task. *Frontiers in Psychology*, *9*, 300.

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and social psychology bulletin*, *34*(2), 196-209.

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

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child development*, *85*(5), 1795-1804.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323-332.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general*, *139*(4), 665.

Kostandyan, M., Bombeke, K., Carsten, T., Krebs, R. M., Notebaert, W., & Boehler, C. N. (2019). Differential effects of sustained and transient effort triggered by reward–A combined EEG and pupillometry study. *Neuropsychologia*, *123*, 116-130.

Kurniawan, I. T., Grueschow, M., & Ruff, C. C. (2021). Anticipatory energization revealed by pupil and brain activity guides human effort-based decision making. *Journal of Neuroscience*.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and brain sciences*, *36*(6), 661-679.

Manohar, S. G., & Husain, M. (2015). Reduced pupillary reward sensitivity in Parkinson's disease. *NPJ Parkinson's disease*, *1*(1), 1-4.

Massar, S. A., Lim, J., Sasmita, K., & Chee, M. W. (2016). Rewards boost sustained attention through higher effort: A value-based decision making approach. *Biological Psychology*, *120*, 21-27.

Moresi, S., Adam, J. J., Rijcken, J., van Gerven, P. W. M., Kuipers, H., & Jolles, J. (2008). Pupil dilation in response preparation. *ScienceDirect, 67,* 124-130. Doi:10.1016/j.ijpsycho.2007.10.011.

Muhammed, K., Manohar, S., Ben Yehuda, M., Chong, T. T. J., Tofaris, G., Lennox, G., Bogdanovic, M., Hu, M., & Husain, M. (2016). Reward sensitivity deficits modulated by dopamine are associated with apathy in Parkinson's disease. *Brain*, *139*(10), 2706-2721.

Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *eLIFE*, *8*, e39404.

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.

Nagase, A. M., Onoda, K., Foo, J. C., Haji, T., Akaishi, R., Yamaguchi, S., ... & Morita, K. (2018). Neural mechanisms for adaptive learned avoidance of mental effort. *Journal of Neuroscience*, *38*(10), 2631-2651.

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.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron, 38(2), 329-337.

Olds, J., & Milner, P. (2020). *3. Positive reinforcement produced by electrical stimulation of Septal area and other regions of rat brain* (pp. 51-66). University of California Press.

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

Oudeyer, P. Y., Gottlieb, J., & Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in Brain Research*, *229*, 257-284.

Sayalı, C., & Badre, D. (2019). Neural systems of cognitive demand avoidance. *Neuropsychologia*, *123*, 41-54.

Sayalı, C., & Badre, D. (2021). Neural systems underlying the learning of cognitive effort costs. *Cognitive, Affective, & Behavioral Neuroscience*, 1-19.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual Review Neuroscience.*, *30*, 259-288.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Ulrich, M., Keller, J., Hoenig, K., Waller, C., & Grön, G. (2014). Neural correlates of experimenttally induced flow experiences. *NeuroImage, 86,* 194-202. Doi:10.1016/j.neuroimage.2013.08.019

Ulrich, M., Keller, J., & Grön, G. (2016). Neural signatures of experimentally induced flow experiences identified in a typical fMRI block design with BOLD imaging. *Social cognitive and affective neuroscience*, *11*(3), 496-507.

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*(1), 1-11.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control. *Psychonomic Bulletin & Review, 26*(6).

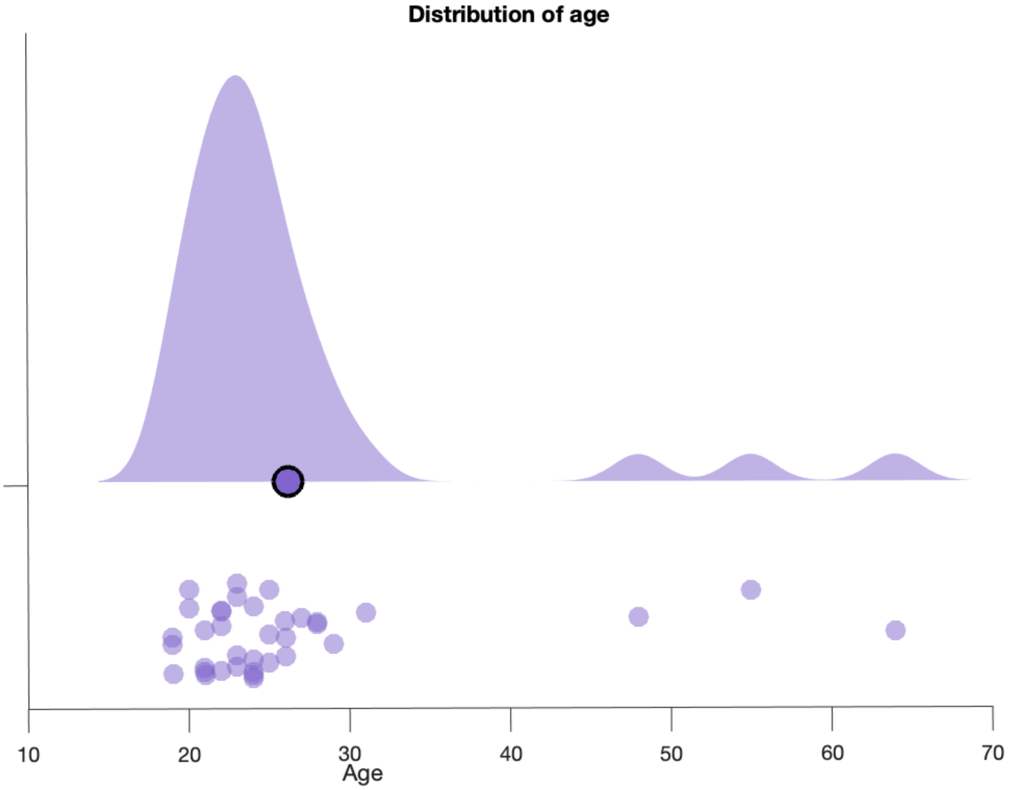
Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PlOS one*, *8*(7), e68210.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and Experiments*, 27-41.

Supplementary Material

Supplementary Methods 1

Distribution of participants' ages was skewed (Supplementary Figure 1). The Shapiro-Wilk statistic associated with participants' ages was W = 0.60, indicating that age distribution significantly deviated from normality ($p < 0.001$).



Supplementary Fig 1. Distribution of participants' ages.

Supplementary Results 1

The effect of Task Difficulty on the subcomponents of Flow inventory

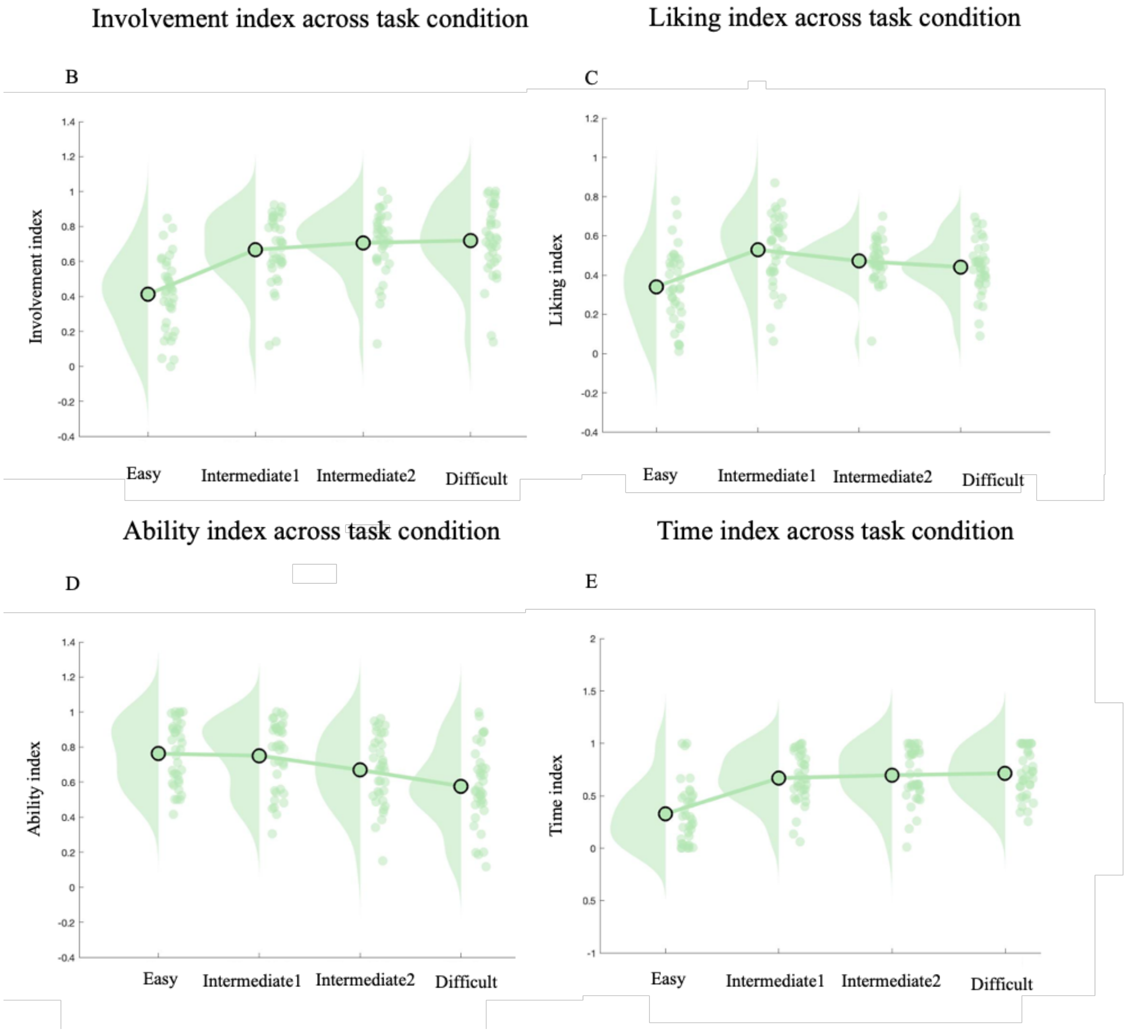
As the flow questionnaire consisted of components of perceived engagement, liking, ability and flow of time, we examined effects of Task Difficulty on each component separately (Supplementary Figure 2).

The results showed a significant effect of Task Difficulty on involvement (*F*(1.479,51,765) = 32.617, *p* < .001), where involvement scores significantly increased with Task Difficulty. Again, subjective engagement scores plateaued at intermediate Task Difficulty. Pairwise t-tests revealed significantly lower scores for easy versus Intermediate1 (*p* < .001), easy versus Intermediate2 (*p* < .001), easy versus Difficult (M = 0.720, SD = 0.035, *p* < .001). No significant differences were found for Intermediate1 versus Intermediate2 (*p* = 0.391), Intermediate1 versus Difficult (*p* = 0.120), nor Intermediate2 versus Difficult (*p* = 1.000).

Liking scores also significantly differed across Task Difficulty (*F*(2.379,83.265) = 11.148, *p* < .001), but this time the effect of Task Difficulty plateaued at Intermediate1 . Specifically, liking scores were significantly lower for Easy versus both Intermediate levels (Intermediate1: *p* < .001; Intermediate2: *p* = .005) and did not differ between Easy vs Difficult (*p* = 0.091), or Intermediate vs Difficult (between intermediate1 and intermediate2 (*p* = 0.339), Intermediate1 and Difficult (*p* = 0.114) nor between Intermediate2 and Difficult (*p* = 0.928)).

Consistent with their own task performance, participants perceived ability at each Task Difficulty level showed a declining trend. Subjective ability scores significantly different across Task Difficulty levels (*F*(2.313,80.955) = 69.095, *p* < .001). Specifically, Easy and Intermediate1 level yielded higher ability compared with Difficult Task Difficulty (Easy: p<0.001; Intermediate1: *p* = 0.005), while other Task Difficulty, including Easy and intermediate levels did not significantly differ from each other (easy vs. intermediate1 (*p* =1.000), easy vs. intermediate2 (*p* = 0.209)), indicating that participants rated their own ability for the easiest and intermediate difficulty levels similarly.

Lastly, perceived time on task showed a significant inclining linear trend across Task Difficulty (*F*(1.428,49.98) = 32.492, *p* < .001). Specifically, both intermediate difficulty levels yielded higher scores compared with Easy (both *p*s < .001), and easy compared with difficult (*p* < .001), while intermediate levels did not significantly differ from each other (*p* = 0.872) nor from Difficult (Intermediate1 vs. Difficult, *p* = 0.643; Intermediate2 vs. Difficult, *p* = 1.000), indicating that subjective time passed quicker with increasing difficulty levels and plateaued at intermediate difficulty.



Supplementary Fig 2. Raincloud plot for B) involvement index, C) liking index, D) ability index, E), and time index across task condition. The dots display each participant's mean score. F) Bar chart showing the mean score of each component for each task condition, error bars indicate standard error.

Supplementary Results 2

The effects of task related parameters on pupil size and the motivational and behavioural relevance of pupil size only in volunteers below 36 years old

We repeated the analysis performed in sections 3.4. - 3.6. only in volunteers that are below 36 years old. We show that the results qualitatively do not change when volunteers above 36 years old (N=3) are excluded from analysis.

Effects of task-related factors on pupil size

Baseline pupil size on the current trial *decreased* with PE (B = -0.56, CI [-0.65, -0.47]), ΔPE (B = -0.12, CI [-0.18, -0.06]), expected accuracy (B = -0.32, CI [-0.47, -0.16]) and Task Difficulty (B = -0.15, CI [-0.18, -0.11]), indicating that smaller baseline pupil size was associated with greater prediction error, learning progress, expected accuracy and task difficulty (Figure **Supplementary Fig 3A**).

On the other hand, pupil size during the cue period (TEPR) was largely predicted by the baseline pupil size. TEPR on the current trial *increased* with smaller baseline pupil size (B = -0.88, CI [-0.91, -0.84]), greater expected accuracy (B = 0.17, CI [0.05, 0.26]) and easier task blocks (Task Difficulty: (B = -0.12, CI [-0.14, -0.10])). The effect of PE and ΔPE were not significant (PE (B = -0.02, CI [-0.07, 0.02]), ΔPE (B = -0.01, CI [-0.04, 0.02]) In other words, phasic pupil response showed an inverse relationship with the baseline pupil size and was greater when the expected accuracy of the upcoming trial was higher (**Supplementary Fig 3B**).

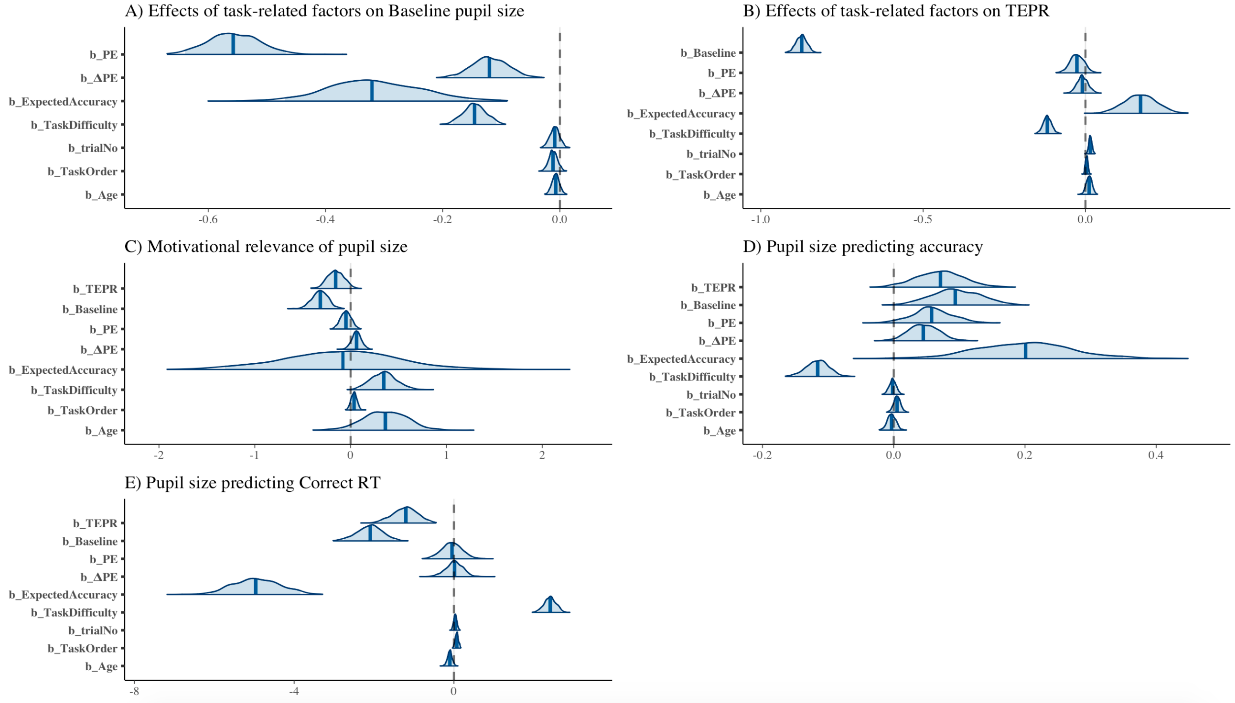
Motivational relevance of pupil size

The model results (**Supplementary Fig 3C**) showed that Baseline pupil size and Task Difficulty but not TEPR, PE, ΔPE were significant predictors of subjective flow scores. Participants reported greater flow scores following trials with smaller baseline pupil size (B = -0.32, CI [-0.48, -0.15]) and greater Task Difficulty (B = 0.35, CI [0.09, 0.62]), indicating that greater subjective engagement was associated with smaller baseline pupil size and greater task difficulty.

Behavioural relevance of pupil size

The model results predicting accuracy (**Supplementary Fig 3D**) showed that the significant predictors of trial-by-trial accuracy were baseline pupil size (B = 0.10, CI [0.02, 0.17]), expected accuracy (B = 0.20, CI [0.05, 0.34]), and Task Difficulty (B = -0.12, CI [-0.15, -0.08]), while the effect of remaining predictors were not significant (TEPR: B = 0.07, CI [-0.00, 0.14]; PE: B = 0.06, CI [0.00, 0.012]. Importantly, baseline pupil size showed a positive correlation with accuracy. As such, accuracy on the upcoming trial was higher when baseline pupil size was greater, while TEPR did not correlate with accuracy.

The model results predicting correct RTs (**Supplementary Fig 3E)** showed that the significant predictors of trial-by-trial RT were TEPR (B = -1.22, CI [-1.82, -0.61]), baseline pupil size (B = -2.11, CI [-2.74, -1.48]), expected accuracy (B = -4.95, CI [-6.05, -3.85]) and Task Difficulty (B = 2.41, CI [2.12, 2.69]). As expected, RTs was longer on trials in more difficult task blocks and when the expected accuracy was lower. Moreover, both smaller baseline pupil size coupled with smaller TEPRs predicted longer RTs.



Supplementary Fig 3. Densities of model parameter estimates. A) Model parameter densities for predicting Baseline pupil size. B) Model parameter densities for predicting TEPR. C) Model parameter densities for predicting flow scores. D) Model parameter densities for predicting accuracy. D) Model parameter densities for predicting Correct RTs.